目　录

第一章 基本概念

§1.1 导 言

（一）什么是数理统计学

关于数理统计学，现在已有了用各种文字出版的大量教科书和专著．在这些著作中，对数理统计学的性质、任务、应用等等，作了不少的论述．应该说，这些问题目前在统计学界并无原则性的分歧．但是，若试图用少量的文字对“数理统计学”这个学科下一个正式的定义，就会碰到不少困难．你很难找到一种说法是完全无懈可击的．况且，任何这样的定义，若不辅之以大量的解释，就无法使人理解．因此，在以下的叙述中，我们将致力于从一些方面把数理统计学的实质说清楚，而不着重于一个形式的定义．

当用观察和实验的方法去研究一个问题时，第一步就是通过观察或试验以收集必要的数据．这些数据受到偶然性即随机性因素的影响．下一步就是对所收集的数据进行分析，以对所研究的问题作出某种形式的结论．在这两个步骤中，都会碰到许多数学问题．为解决这些问题，发展了许多理论和方法．这些就构成数理统计学的内容．故一般地可以说，*数理统计学是数学的一个分支，它的任务是研究怎样用有效的方法去收集和使用带随机性影响的数据*．下面来作些解释．

1．数据必须带有随机性的影响，才能成为数理统计学的研究对象．例如，考虑一个国家的全面人口普查．假定人力物力时间允许我们对国内每一个人的状况进行调查，而这种调查又是准确无误的，则我们可利用普查所得数据，通过既定的方法，把所感兴趣的指标计算出来，例如，男性人口占全体人口的百分之多少，在所作假定之下这是准确无误的．这里不需要用到什么数理统计方

法. 又如要比较两个小麦品种甲、乙谁优(能有更高的产量). 若我们作一个不大现实的假定,即其他条件可以控制得如此严格(且这种条件也是日后大面积推广时所使用的),以致产量完全取决于品种,则我们只须在两块地上把甲、乙各种植一次,就可准确无误地判断其优劣. 在此数理统计方法也没有用武之地. 总之,是否假定数据有随机性,是区别数理统计方法和其他数据处理方法的根本点.

数据的随机性的来源有二: 一是问题中所涉及的研究对象为数很大,我们不可能对之全部加以研究,而只能用"一定的方式"(说详下)挑选其一部分去考察. 例如,一批产品有10,000件,其中含有废品 m 件, m 未知,因而废品率 $p=m/10000$ 也未知. 要确切地知道 p,必须对这10000件逐一加以检验. 这不仅是不经济的,且往往无法做到(如检验是破坏性的). 因此我们只能从其中挑出一部分,例如100件,根据对这100件的检验结果去估计 p. 在这里,随机性的影响就表现在:那100件被挑出是偶然的.

一般,在社会调查性质的问题中,问题的要求规定了调查的范围. 如问题是研究某一地区内以农户为单位的经济状况,则该地区的全体农户都是调查对象. 若这个数目太大,则我们只能挑一部分作实地调查. 这时,所得数据的随机性就来自被挑出的农户的随机性. 对这种数据作分析,就必须使用数理统计方法.

数据随机性的另一种来源是试验的随机误差,这是指那种在试验过程中未加控制、无法控制,甚至不了解的因素所引起的误差. 例如,设反应温度和压力是影响产品质量 Y 的重要因素,我们想通过一定的试验去考察这影响的程度,并挑选一个适当的温度和压力值以供在今后大批生产中使用. 但是, Y 除了与温度、压力有关外,还受到大量其他因素的影响. 例如,每次试验所用原材料略有差异,可能使用不同的仪器设备和操作者等等. 这些因素无法或不便加以完全的控制,而对试验结果(数据)产生随机性的影响. 这就带来一种不确定性. 例如,从试验数据上看,使用温度 t_2 比用 t_1 好. 但这个表现在数据上的优势究竟是本质的——即

有足够的理由可解释为是由于 t_2 确优于 t_1，还是只是随机误差的偶然性表现？这就需要用数理统计的方法去分析．

2．所谓"用有效的方式收集数据"一语中，有效一词该如何解释．归纳起来有两个方面：一是可以建立一个在数学上可以处理并尽可能简单方便的模型来描述所得数据，一是数据中要包含尽可能多的、与所研究的问题有关的信息．

例如，在考察某地区共 10,000 农户的经济状况的问题中，我们前面说挑出 100 户作实际调查．100 这个数字是否恰当？太大了则费用过大，太小了则代表性不够．要决定一个较好的数字，须权衡这两个方面，并用得着统计方法．其次，假定我们选择了 100 这个数字．这 100 户如何挑选？假设你只在该地区最富裕的那部分去挑，这样得到的数据就没有代表性，也谈不上有效了．反之，你如果用一种纯随机化的方法，即设法使这 10000 户中的每一户有同等的机会被挑出，则所得数据就有一定的代表性，我们也不难建立一个简单的模型来描述它．在一些情况下，我们还可以设计出更有效的方法．举一个简单情况．若该地区分成平原和山区两部分，前者较富裕且占全体农户的 70%，则我们可规定，在预定要考察的 100 户中，有 70 户从平原地区挑，30 户从山区挑，而在各自的范围内则用纯随机化的方式挑．直观上我们觉得，这样得到的数据，比在全体 10000 户中用随机化方式挑选得到的数据更有代表性，因而也更"有效"．数理统计的理论证明确是如此．

又如，在产品质量与反应温度和压力的关系的例中，怎样用有效的方式收集数据，问题更多．若可以考虑的温度在 t_1 和 t_2 之间，压力在 p_1 和 p_2 之间．首先，我们当然只能取有限个温度和压力值去做试验．取多少个值好？这里也有与上例中一样的问题：太多了费用太大，太少了不说明问题．在定下了一个数目，例如四个温度值和四个压力值去做试验，则这些值是否均匀地取在相应的区间中好？另外，若把这些值所有可能的搭配都做试验，则至少需做 16 次．也许条件不允许做这么多，而只能做一部分，则这一部分如何挑选？这些问题解决得好，试验数据就有一种平衡或对称的

结构，不仅更富于代表性，且可建立一种简单而便于分析的模型。

用有效的方式收集数据的问题的研究，构成了数理统计学中的两个分支，其一叫**抽样理论**，其二叫**试验设计**，它们分别处理相当于上面讨论过的两个例子中的那种类型的数据收集问题。

3．现在来解释"有效地使用数据"一语的意义．获取数据的目的，是提供与所研究的问题有关的信息．但这种信息并非是一目了然地表现出来，而需要用"有效"的方式去集中、提取，进而利用之以对所研究的问题作出一定的结论．这种"结论"，在统计上叫做"推断"．在§1.3中我们将仔细解释统计推断的意义，这里只指出：所作的推断应是对所提出的问题的一个回答，而不只限于所得数据的范围内．有效地使用数据，就是要使用有效的方法，去集中和提取试验数据中的有关信息，以对所研究的问题作出尽可能精确和可靠的推断．其所以只能做到"尽可能"而非绝对地精确和可靠，是因为数据受到随机性因素的影响．这种影响可以通过统计方法去估计或缩小其干扰作用，但不可能完全消除．

为有效地使用数据以进行统计推断，涉及很多的数学问题．需要建立一定的数学模型，并给定某些准则，才有可能去评价和比较种种统计推断方法的优劣．例如，为估计一物体的重量 a，把它在天平上秤九次，得到数据 $x_1, \cdots, x_9$，它们都受到随机性因素的影响（影响大小反映天平的精密度）．我们可以用这九个值的算术平均 $\bar{x}=\frac{1}{9}(x_1+\cdots+x_9)$ 去估计 a，也可以考虑下述方法：把 $x_1, \cdots, x_9$ 按大小依次排列成 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(9)}$，而取正中间的一个，即 $x_{(5)}$，去估计 a．甚至也可以用两个极端值的平均，即 $w=\frac{1}{2}(x_{(1)}+x_{(9)})$．你可能在直观上会认为：作为 a 的估计，$\bar{x}$ 优于 $x_{(5)}$，而 $x_{(5)}$ 又优于 w．但是为什么？这是不是对？在什么意义下对？在什么条件下对？这些问题就不容易回答．事实上，对这些问题的研究，正是数理统计学的中心内容，要使用大量的数学和概率论的工具．以后我们将看到：在一定的情况（取决于随机性影响的概率结构，即统计模型）和一定的意义（即衡量优越性的指标）之下，上述三个估

计方法中的任一个都可能成为最优的.

4. 最后一点，就是数理统计学只处理在收集和使用带随机性影响的数据中的数学问题，因而是一个数学分支.

一个问题的研究，涉及到问题所在领域的专门知识. 数理统计学不以任何一种专门领域为研究对象：不论你的问题是物理学的、化学的、生物学的或工程技术方面的，只要在安排试验和处理试验数据中涉及到一些一般性的、共同的数学问题，就可以用统计方法. 例如，不论作那种试验，都有一个试验规模的问题，即试验须重复多少次，才能把随机误差的影响控制在必要的限度内. 这是一个与专业知识无关的带共性的问题，一组试验数据，只要对其所受的随机性影响作了明确的规定(如服从正态分布)，则可以用相应的统计方法去分析之，而不管这些数据的实际含义如何. 这种带共性的问题既然从专门的知识领域中超脱出来，就可以用纯数学的方法去研究，这就是数理统计学的对象. 我们这样说，并不意味着一个数理统计学者可以不过问其他专门领域的知识. 相反，如果他要将统计方法用于实际问题，他必须对所论问题的专门知识有一定的了解. 这不仅可以帮助他选定恰当的统计模型和统计方法，而且，用数理统计方法分析随机性数据所得结论的恰当解释，离不开所论问题的专门知识. 例如，数理统计方法对数量遗传学很有用，但一个对遗传学一无所知的统计学家，就难于在这个领域中有所作为.

统计方法的应用很广泛，所以许多学习其他专业的人，都需要一些这方面的知识. 幸好，统计方法的具体使用并不需要很高深的数学知识. 相反，这些方法的理论根据，不具备较多较深的数学知识就说不清楚. 因此，在一些统计方法得到广泛应用的国家，例如在美国，出版了大量专供各领域的应用者使用的著作. 这种著作介绍统计方法及其应用，但不涉及或很少涉及这些方法的理论根据. 这种著作被列入"统计方法"或"应用统计"的范畴内，而只有那种用严格的数学去论证统计方法的理论根据的书，才称为数理统计著作. 这显示在这些国家中，"数理统计学"一词是给以一

种狭义的解释，即只包括统计学中的数学基础部分．在我国，数理统计学一词则是与作为一门社会科学的统计学相对而言的．粗略地说，在我国，数理统计学与西方的统计学相当，而具有较广泛的含义．明白这个差别就可以避免一些误解．

（二）数理统计学的应用

数理统计方法的应用很广泛，几乎在人类活动的一切领域中都能程度不同地找到它的应用．这是因为，实验是科学研究的根本方法，而随机性因素对试验结果的影响是无所不在的．反过来，应用上的需要又是统计方法发展的动力．例如，现代数理统计的奠基人、英国著名学者 R. A. Fisher 和 K. Pearson 在本世纪初期大力从事这方面的研究，就是出于在生物学、数量遗传学、优生学和农业科学方面的需要．

在工农业生产中，一个常见的问题是：有一个（或多个）我们感兴趣的指标，如工业产品的某项质量指标，农业中的单位面积产量等．有一些因素对这个指标可能有影响，例如，工业生产中所用设备、原材料、配方和温度、压力及反应时间等工艺因素．农业生产中所用种子品种，肥料类型和施放数量，以及耕作方法等方面的因素．为了得到最大的经济效益，需要了解这些因素对所感兴趣的指标起影响的具体情况：那些因素是主要的，其影响有多大，因素与指标之间是否可建立某种数量上的联系等等．弄清楚了这些问题，就可确定一组较好的生产条件．这些都要通过做试验，就是把有关因素固定在某些水平上做试验，去观察感兴趣的指标值．试验要经过精心的设计，所得试验结果必然受到大量随机因素的影响，而必须用统计方法分析．因此，随着近几十年来工农业生产的规模愈来愈大，数理统计学在这方面的应用也与日俱增．在历史上说，试验设计的基本思想，以及分析试验数据的一种极重要的方法——方差分析方法，就是 R. A. Fisher 等在 1923～1926 年期间，在进行田间试验中开始发展起来的．Fisher 提出的思想和方法，在四十年代以来经过发展，日益广泛地用于工业生产中．目前

最常用的正交设计，就是一个有代表性的例子.

数理统计方法应用于工业的另一个重要方面是：现代工业生产多有大批量及要求很高的可靠度的特点. 需要在连续生产的过程中进行工序控制，成批的产品在交付使用前要进行验收，这种验收一般不能是全面检验，而只能是抽样验收. 需要根据数理统计学的原理，去制定适合种种要求的抽样验收方案. 还有，一个大型设备往往包含成千上万个元件. 由于元件数目很大，它们的寿命可视为随机的而服从一定的概率分布规律. 整个设备的可靠性，与设备的结构及这种分布规律有关，因而可以用统计方法去估计之. 为解决上面这些问题，发展了一系列的统计方法，目前常提到的"统计质量管理"，就是由这些方法构成的.

统计方法在医药卫生中有广泛的应用. 例如，治疗一种疾病的种种药物和治疗方法的效果，常引用统计资料来说明. 这种材料的可信性，依赖于其数据的取得方法与使用的统计方法. 其他，如分析某种疾病的发生是否与特定因素有关(一个著名的例子是吸烟与患癌症的关系)，关系大小，在污染大气的许多有害成分中，那些成分对人体有何种程度的影响，这类问题常用数理统计方法去研究，取得了不少有用的成果.

数理统计方法在气象预报、地震和地质探矿等方面有一些应用. 在这类领域中，人们对事物的规律性认识尚不充分，使用统计分析方法可能有助于获得一些对潜在的规律性的认识，而用以指导人们的行动. 不过，在人们对事物的规律性认识很不充分的情况下，一些起比较大的作用的系统性因素，只好当作随机性因素来处理，这样，统计分析的精度或可靠性就较差.

自然科学的任务是揭示自然界的规律性. 一般是先根据若干观察或试验资料提出某种初步理论或假说，然后再从种种途径通过实验去验证之. 在这里统计方法起相当的作用. 一个好的统计方法有助于提取观察和试验数据中带根本性的东西，因而有助于提出较正确的理论或假说. 在有了一定的理论或假说后，统计方法可以指导学者如何去安排进一步的观察或试验，以使所得数据

更有助于判定理论或假说是否正确．数理统计学也提供了一些理论上健全的方法，以估量观察或试验数据与理论的符合程度如何。一个著名的例子是遗传学中的 Mendal 定律．这个根据观察资料提出的定律，经历了严格的统计检验．数量遗传学的基本定律——Hardy-Weinberg 平衡定律，也是属于这种性质．

数理统计方法在社会、经济领域中也有很多应用．在某些国家中，统计方法在这些方面的应用，比其在自然科学和技术领域中的应用更为显著．统计方法在社会领域中的一项重要应用是抽样调查．经验证明，经过精心设计和组织的抽样调查，其效果可以达到以至超过全面调查的水平．另外，对社会现象的研究有向定量化发展的趋势，例如人口学，确定一个合适的人口发展动态模型，需要掌握大量的观察资料，并使用包括统计方法在内的一些科学分析方法．在经济科学中，定量化的趋势比其他社会科学部门更早更深．早在本世纪二、三十年代，时间序列的统计分析方法就用于市场预测．现在，一系列的统计方法，从简单的到很艰深的，都在数量经济学和数理经济学中找到了应用．

（三）简单历史

下面我们简单地介绍一下数理统计学这门学科的简单历史．这必然是极为“粗线条”的，因为叙述一门学科的历史，离不开这门学科的具体内容．另外，关于数理统计学的早期发展情况，学者们很少论述，可资征引的文献很少．

数理统计学是一门较年青的学科，它主要的发展是从本世纪初开始．在早期发展中，起领导作用的是以 R. A. Fisher 和 K. Pearson 为首的英国学派．特别是 Fisher，在本学科的发展中起了独特的作用．目前许多常用的统计方法以及教科书中的内容，都与他的名字有关．其他一些著名的学者，如 W. S. Gosset (Student)、J. Neyman、E. S. Pearson (K. Pearson 的儿子)、A. Wald 以及我国的许宝騄教授等，都作出了根本性的贡献．他们的工作奠定了许多统计分支的基础，提出了一系列有重要应用

价值的统计方法，和一系列的基本概念和重要理论问题．有一种意见认为，瑞典统计学家 H. Cramer 在 1946 年发表的著作《Mathematical Methods of Statistics》标志了这门学科达到成熟的地步．这样说也许并不过分，因为虽则在此以前已出现了一些重要的统计著作，特别是 R. A. Fisher 的《Experimental Design》和《Statistical Methods for Research Workers》．但第一次用严整的数学方法总结到那时为止数理统计学的主要成就的，还是要推 Cramer 的上述著作．

收集和记录种种数据的活动，在人类历史上很久远．翻开我国的二十四史，可以看到上面有很多关于钱粮人口及地震洪水等自然灾害的记录。在西方，Statistics(统计学)一词源出于 State(国家)，意指国家收集的国情资料．也有不少人为研究特定问题而进行观察试验，收集资料．但这些情况，终究还不能认为是数理统计这门学科已经成立的标志．因为有许多工作，只停留在收集数据或对之进行一些简单的加工整理．即使有时也作出了某种越出已有数据范围之外的推断，也只是基于一种朴素的直观想法，而未能把问题模型化使之带有普遍意义，更谈不上建立必要的基本概念和理论了．这种情况延续了许多年，这是因为，没有一定的数学工具特别是概率论的发展，无法建立现代意义下的数理统计学．也因为应用方面的要求还没有达到那么迫切，足以构成一股强大的推动力．到上世纪后半期直至本世纪初，情况才起了较大的变化．是否可以举出某一个时间或某一部著作，足以作为数理统计学正式诞生的标志？学者们提出过一些意见，但尚无定论．有的认为这个时间"不早于 1850 年"，有的将它定在 Fisher 诞生的那一年——1890 年．有的认为本世纪初 K. Pearson 关于 χ^2 统计量的极限分布的论文可以作为一个标志，也有人认为，直到 1922 年 Fisher 关于统计学的数学基础的那篇著名的论文发表，数理统计学才正式诞生．这个时间可能失之过晚，不过，Fisher 的这篇论文首次概括了统计理论的现状和存在的问题，并提出了数理统计学的三个任务．文中的观点的主要部分到现在仍基本有效．因之，

Fisher 这项工作无疑是数理统计学建立过程中的一个里程碑.

综合以上所述,我们可否试探性地下这样一个粗线条的结论:收集和整理乃至使用观测和试验数据的工作由来已久,这类活动对于数理统计这门学科的产生,可算是一个源头. 上世纪特别是上世纪后半期以来发展速度加快,且有了质的变化. 在上世纪末期到本世纪初期这一段,出现了一系列的重要工作. 无论如何,至迟到本世纪二十年代,这门学科已稳稳地站住了跟脚. 本世纪前四十多年有了迅速而全面的发展,到四十年代时,已形成为一个成熟的数学分支.

本世纪前四十多年是数理统计学辉煌发展的时期. 但战后以来这几十年,数理统计学的发展也很显著. 许多在战前开始成形的统计分支,在战后得到纵深的发展. 数学上的深度比以前大大加强了. 也出现了若干带根本性的新发展,如 Wald 的统计判决理论与 Bayes 学派的兴起(均见第五章),在数理统计的应用方面,也给人深刻的印象. 这不仅是由于战后工农业和科技等方面迅速发展所提出的要求,也由于电子计算机这一有力工具的出现. 许多统计方法的实施都涉及很大量的计算. 在电子计算机得到广泛应用以前,这些方法的威力就难于发挥. 例如在五十年代,当电子计算机还很稀少时,用电动计算机计算一个包含十余个自变量的线性回归,得用几十个人花成月的时间,现在在大型计算机上,所花时间则只以秒为单位计. 以此之故,有些在战前就已得到充分发展的统计方法,真正在应用上发挥作用还是在战后.

§1.2 样本和样本分布

在上节中,我们对"什么是数理统计学"这个问题,作了一番定性式的描述. 其所以是定性的,因为我们没有引进必要的数学概念及使用严密的数学语言. 在一个意义上说,本章其余这几节是继续上节的工作,对"什么是数理统计学"这个问题给以更细致而严密的回答. 在这个过程中,也就自然地引进了一些重要的基本

概念。

(一) 样本

通过观察或试验而得到的数据，称为样本，又称样品，子样．如在同一架天平上将一个物件秤 n 次，得到数据 $X_1, \cdots, X_n$，则它们的全体，即 $X=(X_1, \cdots, X_n)$，就称为样本．n 称为样本大小，也有称为样本容量或样本含量的．数据 $X_1, \cdots, X_n$ 中的每一个，例如 X_i，也称作样本．这不致引起混淆．

在很多情况下，样本 X_i 是数量性的，即取实数为值．当然，也可能只取某些特殊的实数值，例如非负整数．在另一些情况下，样本 X_i 是"属性"数据．例如，产品可分为甲、乙、丙三级，一个人可以是有病或无病．这种属性数据可以数量化．例如，甲、乙、丙三个等级可以分别用 1、2、3 这三个数字代替．这种数量化并非总是必要．

经常，每次观察或试验所记录的，不止一个实数．例如，在一大群人中抽取 n 个，每个测得其身高体重，为 $(X_1, Y_1), \cdots, (X_n, Y_n)$．则我们有一组大小为 n 的**二维样本**．类似地有**多维样本**．多维样本中的任一个成分称为样本的**分量**．如在本例中，身高 X_1 是样本 (X_1, Y_1) 的一个分量．

从实用者的眼光看，样本就是一批已知的数字．但我们不能忘记：数理统计学的对象是受到随机性影响的数据．用概率论的语言说，样本是随机变量．表现为已知数字的具体样本，则是这随机变量的观测值．样本的这种**二重性**虽然是一件平凡的事情，但有很大的重要性．对理论工作者而言，他更多注意到样本是随机变量这一点．因为他所发展的统计方法应有一定的普遍性，不止是可用于某些具体样本值．反之，对应用工作者而言，他们虽则习惯了把样本看成具体数字，但仍然不能忽视"样本是随机变量"这个背景．不然的话，样本就不过是一堆杂乱无章的、毫无规律性可言的数字，没法进行任何统计处理．

样本 $X=(X_1, \cdots, X_n)$ 可能取的值的全体的集，称为**样本空**

间．拿上面秤东西的例子说，如果我们能定出两个实数 $a<b$，使 1°．此物件在天平上秤量的结果，不会小于 a 也不会大于 b．2°．区间 $[a, b]$ 内的每个值都可能取．则样本空间当为 n 维欧氏空间 R^n 中的子集

$$\mathscr{X}=\{(x_1, \cdots, x_n): a\leqslant x_i\leqslant b, i=1, \cdots, n\}.$$

但这一来问题就很麻烦：因为即使在本例这样简单的场合下，a、b 的确切值也不易定出．因此，我们把样本空间的定义修改为：若某个集合 $\mathscr{X}$ 包含了一切可能的样本值，则 $\mathscr{X}$ 称为样本空间．如在上例中，我们可以万无一失地把样本空间取为 $\{(x_1, \cdots, x_n): 0\leqslant x_i<\infty, i=1, \cdots, n\}$，甚至取为整个 R^n．这样一来，样本空间中难免会包含一些多余点．这没有什么关系，因下面我们将看到：重要的是样本的分布．如果样本空间中，实际上不可能出现的点所占概率为 0 或很小，则对问题的实质毫无影响或影响很小．

（二）样本分布

样本既然是随机变量，就有一定的概率分布，这个概率分布就叫做**样本分布**．

我们曾多次指出：样本，也就是观察或试验数据，是受到随机性的影响的．但是这种影响的具体方式如何，则取决于所观察的指标的性质、观察手段和方法等．所有这些，都总结到一个东西——样本分布中去．就是说，样本分布是样本所受随机性影响的最完整的描述．

因此，要决定样本的分布，就需要根据所观察的指标的具体性质（这往往涉及有关的专业知识），以及对抽样的方式或试验进行的方式的了解．即使这样，一般所了解的情况仍不足以完全决定样本的分布，而必须加进一定程度的假设成分．我们来看一些例子．

例 1.1 一大批产品共有 N 个，其中有废品 M 个．N 已知而 M 未知．现在要从中抽出 n 个加以检验，用以估计 M 或废品率 $p=M/N$．抽样的方式是这样的：第一次抽一个时，N 个产品

中的每一个有同等机会$\left(\text{即}\ \frac{1}{N}\right)$被抽出. 抽出后, 还剩下 $N-1$ 个, 第二次抽一个时, 剩下这 $N-1$ 个的每一个有同等机会被抽出, 以此类推, 直到抽完 n 个为止.

在本例中我们有一个属性指标. 先将其数量化, 让废品对应 1, 合格品对应 0. 这样, 样本 $X_1, \cdots, X_n$ 中的每一个都只能取 0, 1 为值. 给定一组 $x_1, \cdots, x_n$, 每个 x_i 为 0 或 1. 我们来计算概率 $P(X_1=x_1, \cdots, X_n=x_n)$. 为便于讨论, 设 $x_1=1$, $x_2=0$, $x_3=1$. 为要 $X_1=1$, 第一个必须抽得废品. 按"每一个有同等机会被抽出"的规定, 这个事件的概率为 $\frac{M}{N}$. 到第二次抽时, 还有 $N-1$ 个产品, 其中合格品有 $N-M$ 个. 故 $X_2=0$, 即第二次抽得合格品, 概率为 $\frac{N-M}{N-1}$[1]. 同理推得 $X_3=1$ 的概率为 $\frac{M-1}{N-2}$. 这样下去就不难算出所要的结果. 易验证:

$$P(X_1=x_1, \cdots, X_n=x_n)=\frac{M}{N}\frac{(M-1)}{(N-1)}\cdots$$

$$\frac{(M-a+1)}{(N-a+1)}\frac{(N-M)}{(N-a)}\cdots\frac{(N-M-n+a+1)}{(N-n+1)},$$

$$\text{当 } x_1, \cdots, x_n \text{ 都为 0 或 1}, \sum_{i=1}^{n} x_i=a \text{ (其他情况为 0)}. \quad (1.1)$$

这就是本例的样本分布.

在有限总体, 即总体中只包含有限个个体的情况下, 这种随机抽样——即使得总体中每一个体有同等机会被抽出的抽样方式, 是一种最基本的抽样方式. 更确切地说, 设总体中有 N 个个体而预定抽出 n 个, 则随机抽样的意义是, 使总体中任何特定的 n 个有 $\binom{N}{n}^{-1}$ 的机会被抽出. 在具体工作中这一点不总是容易实现的. 例如要在一县的农户中随机抽 n 个, 则可能需要制作 N 张卡片, 把全部农户编号填在各卡片上, 再将卡片彻底混乱后抽出 n 张. 这

1) 这个概率严格地应理解为条件概率 $P(X_2=0|X_1=1)$, 因为它是在 $X_1=1$ 的条件下算出的. 以下类似.

还只涉及工作上的繁琐问题，有时更存在实质性的困难．比方说，要研究经常打扰其睡眠对一个人(在一定范围内)在情绪上的影响．这种试验必须志愿参加，这就带来一种非随机性．或者，试验者可以从他认识的人里面去征求试验对象，这对他研究的总体而言也是非随机的．何况，受试验者对试验目的的了解，可能会使得试验结果受到影响．在这种及类似情况下，真正的随机抽样很难实现．而研究者必须正确估量：他抽样的那个实际(而非名义)总体究竟是什么．

例 1.2 仍考虑上例，但修改抽样方式如下：每次抽一个，记录结果后，将其放回去，再抽第二个……，直到抽出 n 个为止．且在每次抽取时，N 个产品中每一个有同等机会被抽出．

仍以 $X_1, \cdots, X_n$ 记样本．在此，不论前 $i-1$ 次抽取的结果如何，到第 i 次抽取时，总是有 N 个产品，其中废品 M 个．故 $\{X_i=x_i\}$，$i=1, \cdots, n$ 这 n 个事件独立，且 $P(X_i=x_i)=\dfrac{M}{N}$ 或 $\dfrac{N-M}{N}$，视 $x_i=1$ 或 0 而定．于是得到

$$P(X_1=x_1, \cdots, X_n=x_n)=\left(\frac{M}{N}\right)^a\left(1-\frac{M}{N}\right)^{n-a},$$

$$\text{当 } x_1, \cdots, x_n \text{ 都为 } 0 \text{ 或 } 1, \sum_{i=1}^{n} x_i=a; \tag{1.2}$$

$$\text{其他情况}, \sum_{i=1}^{n} x_i=0.$$

例 1.1 和例 1.2 中的抽样方式分别称为“无放回的”和“有放回的”．这个差别对样本分布有影响．也可以说成：在这两种不同的抽样方式之下，随机性的影响也有所不同．(1.2)比(1.1)简单得多，这是由于，在例 1.2 中样本 $X_1, \cdots, X_n$ 是独立同分布的，而在例 1.1 中样本 $X_1, \cdots, X_n$ 不独立．当 N/n 很大时，(1.1)与(1.2)相差很小，因而在这种情况下，可以近似地把不放回的抽样当作有放回的抽样来处理．

例 1.3 某地区一共有 N 个农户，N 已知．从这 N 个农户中抽取 n 户去考察其经济情况．这里抽样也有不放回和放回两

种，从实际工作的角度看抽样必然是不放回的，但如 N/n 很大，可近似地看成是有放回的.

确定一个指标，例如年纯收入. 把 N 户按任何指定的方式编号，从 1 开始至 N，以 a_i 记第 i 户的年纯收入. 抽出的 n 户的年纯收入，即样本，记为 $X_1, \cdots, X_n$. 假定抽取是随机的，即在有放回时，每次抽取抽得任一户的概率为 $\frac{1}{N}$，而在无放回时，则抽得还剩下的户中任一户的概率，等于剩余下的户数的倒数. 这样就不难求得 $X_1, \cdots, X_n$ 的分布. 暂设 $a_1, \cdots, a_N$ 彼此不同，则结果为：

1. 无放回时

$$P(X_1=a_{i_1}, \cdots, X_n=a_{i_n})=\frac{1}{N(N-1)\cdots(N-n+1)}, \quad (1.3)$$

此处 $i_1, \cdots, i_n$ 为彼此不相等的，不超过 N 的自然数.

2. 有放回时

$$P(X_1=a_{i_1}, \cdots, X_n=a_{i_n})=\left(\frac{1}{N}\right)^n. \quad (1.4)$$

若 $a_1, \cdots, a_N$ 中有相同的，则以 $b_1, \cdots, b_l$ 记其中的不同值，且设 $a_1, \cdots, a_N$ 中有 N_1 个 b_1，N_2 个 b_2，$\cdots$，N_l 个 b_l，则也易算出

1. 无放回时

$$P(X_1=x_1, \cdots, X_n=X_n)=\prod_{i=1}^{l}[N_i(N_i-1)\cdots(N_i-n_i+1)]/[N(N-1)\cdots(N-n+1)]. \quad (1.5)$$

2. 有放回时

$$P(X_1=x_1, \cdots, X_n=x_n)=N_1^{n_1}N_2^{n_2}\cdots N_l^{n_l}/N^n, \quad (1.6)$$

其中 $x_1, \cdots, x_n$ 有 n_1 个为 b_1，n_2 个为 b_2，$\cdots$，n_l 个为 b_l.

在以上这种性质的几个例子中，样本分布在下述意义上说是确切的. 就是说，它是根据严格的概率计算得到，不含有什么假定成分——如果有的话，那只在于：我们要假定“每一个体有同等机会被抽出”这一点，是可以严格实现的. 在具体实施中，可以采用随机数表之类的工具去做. 虽则完全严格地实现“同等机会”也许不可能，但经过精心计划和实施的抽样程序，可以“基本上”实现这

一点，其偏差从应用角度看可以忽略．除了这几个例子所代表的那些情况外，样本分布的确定都带有不同程度的假设成分，有时甚至就是一种假设．当然，这种假设往往有一定的理论或经验的依据，但数学上的简单性往往也是一个因素．看以下的例子．

例 1.4 为估计一物件的重量 a，用一架天平将它重复秤量 n 次（也可以用不同的天平，但精度相似），结果记为 $X_1, \cdots, X_n$，这就是样本．

要定出样本 $X_1, \cdots, X_n$ 的分布，就没有前几个例子那种简单的算法，而只能是一系列假定的结果．首先，假定各次秤量是"独立"地进行的．这可理解为：某次秤量的误差大小，不受其他次秤量结果的影响．这样，$X_1, \cdots, X_n$ 是相互独立的随机变量．其次，假定各次秤量是在"同样条件下"进行的．这可理解为：每次用同一架天平，或各次使用的天平在精度上可认为一样；每次秤量由同一个人操作，或操作者虽不同，但他们的技术水平可认为相同；每次操作时，操作者都是专心贯注，且周围环境条件也可认为一样，等等．这些条件在实际中也许不能彻底做到，但我们假定偏离很小以致其影响可忽略不计．在这个假定下，$X_1, \cdots, X_n$ 中每一个有同样的分布．因此，样本 $X_1, \cdots, X_n$ 是 n 个独立同分布的随机变量．

这样一来，为确定$(X_1, \cdots, X_n)$的概率分布，只须给出其中一个，例如 X_1 的概率分布．在此就要考虑到秤量误差的本性．这种误差一般是由大量的、彼此独立地起作用的随机性误差迭加而成，每一个起的作用不大．由概率论中的中心极限定理可知，这种误差近似地服从正态分布．再假定天平没有系统性误差，则可进一步假定此正态分布有均值 0．于是，可以把 X_1（它等于物件重量 a 加上秤量误差）的概率分布定为正态 $N(a, \sigma^2)$．方差 σ^2 反映天平的精度，而样本$(X_1, \cdots, X_n)$的分布有概率密度

$$f(x_1, \cdots, x_n; a, \sigma) = (\sqrt{2\pi}\sigma)^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2\right]. \tag{1.7}$$

在本例中，为得出分布(1.7)，所引入的假定有两种类型：一种是导出“$X_1, \cdots, X_n$ 为独立同分布”的假定，它可以说与前几个例中，“每一个体有同等机会被抽出”的假定相当。另一种是 X_1 为正态 $N(a, \sigma^2)$ 的假定。这一点依据问题的性质，概率论的理论，以往的经验等等，而终不免是一种多少带有任意性的假定。这就是本例与前面三个例子不同的地方。

例 1.5 某工厂生产一种电子元件，如晶体管。由于在大批生产中种种随机性因素的干扰，所生产的晶体管寿命不同。我们想抽取 n 个作试验，用以估计其平均寿命。

设用一定的方法抽取了 n 个元件，将其在正常(或指定)的条件下使用，直到全部 n 个都失效为止，测得这 n 个的寿命分别为 $X_1, \cdots, X_n$，这就是样本。

要确定$(X_1, \cdots, X_n)$的分布，要作一系列的假定。首先，假定这工厂的生产量很大，生产条件在所考察的那段时间内基本上是稳定的。又这 n 件产品的抽取，尽量照顾到不是在同一天，同一个班次或同一个人的产品，而是在时间、空间等方面有足够的“跨度”，则有理由假定 $X_1, \cdots, X_n$ 是独立且同分布的随机变量。这一点与例 1.4 相似。

为要确定 X_1 的分布，则须作进一步的假定。例如，假定寿命有无后效性，即在元件已使用了长为 t 的一段时间尚未失效的条件下，元件至少尚能使用一段时间 s 的概率，不依赖于 t。又设当元件在时刻 t 尚未失效时，它在 $[t, t+\Delta t]$ 的时间段内失效的概率，有 $\lambda\Delta t+0(\Delta t)$ 的形式(其中 $\lim\limits_{\Delta t\to 0} 0(\Delta t)/\Delta t=0$，$\lambda>0$ 为一常数)，则在概率论中证明 $P(X_1<x)=1-e^{-\lambda x}$，当 $x>0$ (当 $x\leqslant 0$ 时，此概率当然为 0)，即有指数分布。也可以写成概率密度的形式：X_1 有概率密度

$$
\begin{aligned}
f(x, \lambda) &= \lambda e^{-\lambda x}, \text{当 } x>0; \\
&= 0 \quad\quad\;, \text{当 } x\leqslant 0.
\end{aligned} \tag{1.8}
$$

在本例中，样本$(X_1, \cdots, X_n)$的分布是用严格的概率论方法

推出的. 但是,推导所根据的假定,在实际问题中至多只能是近似地成立,因此,对样本分布的规定也多少只是一种假定. 当进行理论研究时,我们是以这种假定严格满足为出发点. 因此,建立在这一点上的理论和方法是否在一特定情况下可用,就要具体考察该特定情况下的条件.

例 1.6 若样本是多维的(见本节(一)),以上诸例的讨论也适用. 例如,从一大群人中抽出 n 个,每个测出其身高和体重,得 (X_1Y_1), $\cdots$, (X_n, Y_n). 这就是样本. 在一定条件下,可假定它们独立同分布,且(X_1, Y_1)有二维正态分布 $N(a, b, \sigma_1^2, \sigma_2^2, \rho)$. 于是样本$((X_1, Y_1), \cdots, (X_n, Y_n))$有概率密度

$$f(x_1, y_1, \cdots, x_n, y_n; a, b, \sigma_1^2, \sigma_2^2, \rho)$$
$$=\prod_{i=1}^{n}\left[\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\left(\frac{(x_i-a)^2}{\sigma_1^2}-\frac{2r(x_i-a)(y_i-b)}{\sigma_1\sigma_2}+\frac{(y_i-b)^2}{\sigma_2^2}\right)\right)\right]. \tag{1.9}$$

(三) 统计模型

在 § 1.1 中我们多次提到"统计模型"这个词. 现在很容易解释其确切含义: 所谓一个问题的**统计模型**,就是指研究该问题时所抽样本的分布,也常称为**概率模型**或**数学模型**. 由于模型只取决于分布,故常把分布的名称作为模型的名称. 这样,例 1.4 和例 1.6 中的模型可称为正态分布模型或正态模型,例 1.5 中的模型可称为指数分布模型,等等. 就上面提出的模型定义,有几点可以解释一下.

1. 在所作的定义下,模型是对确定的样本而言. 就是说,只有在明确了样本的产生方法,并辅之以必要的假定,才能定下模型. 照这样看,单说是"问题的模型"还不很合适. 因为提出问题并非必须给定样本不可,好比从 10,000 农户中抽出 100 户去考察其经济情况那个例子(见 § 1.1,(一),2),问题可以是估计平均年纯收入. 这问题是明确的,如何抽样,可以有种种方法. 我们这样托

模型和样本紧密联系是方便的：统计分析的依据是样本．从统计[1]上说，只有规定了样本的分布，问题才算明确了．

2．也可以从另一面来说：什么时候给出了样本的分布，就有了确定的统计模型，而不在乎问题怎么提，或换句话说，在一个模型之下可提出很多统计问题．拿例 1.4 而言，有了样本 $X_1, \cdots, X_n$，并规定其分布为(1.7)后，就有了一个统计模型．在这个模型下可提出一些问题，例如估计物件的重量 a．也可以设想：我们的目的不在于估计 a，而在于考察这天平的精度如何．在统计上可以把这个问题提成估计 σ^2．总之，以上两点就是说：统计模型就是样本分布，不管抽样的目的何在．

把统计模型定义为样本分布，包含了一些较深刻的意思，我们来仔细申述一下．第一，这表明，数理统计学是通过研究样本的分布解决所提出的问题．或换句话说，从统计学的观点看，样本中的信息，全都在其分布中．因此，这种信息对要解决的问题目，用处多大，就要看样本分布与问题是否有关，关系如何．4 来说，问题是估计物件重量 a，而样本的分布(1.7)确与 a 有关．在例 1.5 中，元件的平均寿命是分布(1.8)的均值 $\frac{1}{\lambda}$．这个值取决于分布(1.8)．在这两例中，样本中都包含了有关问题的有用信息．以后将看到，在包括这两个例子在内的一些情况中，我们甚至可以在一定的意义下定量地衡量这信息量的大小．这样，在 §1.1 中多次提到的“有效地收集数据”一语，也就可给以更确切的解释：“有效”是指所得样本的样本分布中，能包含尽可能多的、与问题有关的信息．

第二，由于模型就是样本分布，故很多性质不一样的问题，可归入到同一种模型下．例如，所有涉及到量测误差的问题，只要例 1.4 中申述的、假定误差服从正态分布的理由成立，则都可以用模型(1.7)．只要把这个模型的统计问题研究清楚了，就可用于解决许多专业部门的问题．这就是数学的抽象．以此之故，可以说数

1) 更确切地说，应是指统计推断而言，因为统计学也处理收集样本的问题．

理统计学的任务，就是研究种种统计模型中所能提出的种种统计问题．它形式上可以从任何具体的专业领域中超脱出来，而[illegible]纯数学的．当然，选定有意义的统计模型，提出有意义的统计[illegible]题，对结果作恰当的解释和利用，这些都离不开实际背景．

第三，由这个提法看出数理统计学和概率论的极密切的关系．研究数理统计学主要的工具是概率论．在此有一个问题需要明确：既然统计模型就是概率分布，而概率分布又是概率论的主要研究对象，那么，数理统计学和概率论还有何区别？可否把数理统计学看成概率论的一部分？我们说不能，尽管二者关系如此密切，它们还是两个独立的、平行的学科，主要原因是所研究的问题的性质不同．拿例 1.4 来说，正态分布在概率论中有很深入的研究，但目的只在于弄清"正态分布有何数学性质"．而数理统计学所关心的，则是如何用样本去推断这分布中未知的成分，即 a 和 σ^2 [illegible]题．这有点象解析几何与代数的关系：研究解析几何使[illegible]具，但解析几何因有其独特问题而自成一学科．更具体一[illegible]论中研究了许多分布的性质，这些分布都可作为统计模型[illegible]论所提供的有关这些分布的知识，可用于解决关于这些分布[illegible]计问题．以此之故有一种说法：概率论是统计的基础，统计是概率论的一种应用．这个说法基本上正确地概括了二者的关系．

（四）总体

总体又称**母体**．在统计上，这个词常理解为"研究的问题所涉及的对象的全体的集合"．总体中的每个成员则称为**个体**或**单元**．从总体中按一定的规则抽出一些个体的行动，称为**抽样**．所抽得的个体称为**样本**．

拿例 1.1 来说，因问题只在估计这批产品的废品率（而不及其他），故总体就是这批产品，抽出的 n 个产品就是样本．这里与本节（一）中所述有一点小小的区别．按这里的说法，样本是产品本身，而按（一）的说法，则是它们的特征（数字 1 或 0）．在问题已明确时，这一差别自无关宏旨．同样，在例 1.3 中，总体是该地区全

部 N 个农户的集，每一农户是个体，而样本包含 n 个农户．在类于此类例子的问题中，总体是由有限个“看得见、摸得着”的人或构成，一切都比较好理解．

再看例 1.4．在这里总体该如何理解？这就不象以上诸例那么清楚．实际上，在此例及类似的情况下，总体并非现实存在的对象的集，而只是我们头脑中的抽象：我们把总体理解为“一切可能出现的秤量结果的集”．如果愿意的话，你可以这样想：若在尽可能同样的条件下把这物件无尽期地秤量下去，则会得到无数的秤量结果．把它们一一记录在一本有无穷多页的书上，这本书就是问题的总体．简化一些，我们定出一个区间 $[a, b]$，它包含了一切可能的秤量结果．就以这个区间作为总体，比方说，你可以万无一失地取 $[a, b]$ 为 $[0, \infty)$，以至 $(-\infty, \infty)$．这区间内每一个数都是布么个体，而样本仍可理解为总体的一部分．但是，按这样方式定所包含就失掉了其特性：只要所考察的指标取实数值，我们总可是否有取为 $(-\infty, \infty)$．我们还不如就从样本分布(1.7)出发，拿例 1 什么总体．

是，在统计著作中常引进**总体分布**这一概念，这个概念的引入，使得在某些重要情况下，从总体出发比从样本出发有其方便之处．我们把总体分布定义为当样本大小为 1 时的样本分布．就是说，若只观察一次或只做一次试验，则所得结果，即 X_1 的概率分布，就是总体分布．拿例 1.4 和例 1.5 来说，总体分布分别是正态 $N(a, \sigma^2)$ 及指数分布(1.8)．如果这样定义总体分布并以 F 来记它，则当有一个抽自这总体的大小为 n 的独立同分布样本 $X_1, \cdots, X_n$ 时，可立即写出样本分布为 $F(x_1)F(x_2)\cdots F(x_n)$．这样，就可以把统计模型定义为总体分布 F，这样定义避免了把模型定义与样本大小 n 关联起来，而显得较简单．由于独立同分布样本是统计中最常考虑的对象，总体分布是一个有用的概念．

引进总体分布这个概念，也解决了上面提出过的一点，即如果单看指标值的集，则总体都可取为 $(-\infty, \infty)$ 而失掉其特性．现在可以说：总体的特性由总体分布来刻划．因之在统计上，常把总

体和总体分布视为同义语，也根据总体分布的类型来称呼总体．如例 1.4 中总体可称为“正态总体”，或“总体 $N(a, \sigma^2)$”． 当总体分布为 F 而 $X_1, \cdots, X_n$ 为独立同分布样本时，常称 $X_1, \cdots, X_n$ 是从总体 F 中抽出的简单随机样本或独立随机样本，并记为

$$X_1, \cdots, X_n \sim F. \tag{1.10}$$

若分布 F 有密度 f，则也常记为

$$X_1, \cdots, X_n \sim f. \tag{1.11}$$

另一种在统计上常用的说法如下：我们引进一个抽象的记号，例如 X，来代表所考察的指标． 如在例 1.4 和 1.5 中，记号 X 分别代表“秤量结果”与“元件寿命”． X 不是样本，因为它并不是由实际观察或试验产生，而只是我们所考察的量的一个记号．我们把 X 看成一个随机变量，其分布就是总体分布 F．这样，我们可以把总体就看成是变量 X，而样本 $X_1, \cdots, X_n$ 是 X 的观察值．记为

$$X_1, \cdots, X_n \sim X. \tag{1.12}$$

(1.10)～(1.12) 这三个记法表示同一个意思：样本 $X_1, \cdots, X_n$ 独立同分布，每个样本(如 X_1)的分布与 X 的分布同，即有分布 F，或密度 f．

在许多情况下，整个样本并非独立同分布，但可以自然地分成几部分，每部分都是独立同分布．看下面的例子．

例 1.7 两个物件的重量 a、b 未知，为比较 a 和 b，我们在同一架天平上把它们分别秤 m 和 n 次，结果分别记为 $X_1, \cdots, X_m$ 和 $Y_1, \cdots, Y_n$.

全部样本是$(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$．假定例 1.4 中陈述的那些条件适合，则有理由认为 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ 这 $m+n$ 个变量相互独立，X_i 服从正态分布 $N(a, \sigma^2)$ 而 Y_j 服从正态分布$N(b, \sigma^2)$．在此 $X_1, \cdots, X_m$ 同分布，$Y_1, \cdots, Y_n$ 同分布，但全体样本不同分布．象这类的例子，我们不如采取这样的看法：有两个总体，分别以 X, Y 记之，总体分布分别是 $N(a, \sigma^2)$ 和 $N(b, \sigma^2)$．样本也分为两组，即$(X_1, \cdots, X_m)$和$(Y_1, \cdots, Y_n)$，它们分别抽自总体 X 和 Y．可记为：

$$X_1, \cdots, X_m \sim X; Y_1, \cdots, Y_n \sim Y. \qquad (1.13)$$

在统计上称这种问题为"两样本问题"(当然,称之为两总体问题也未尝不可). 类似地可定义多样本问题.

§1.3 统计推断

在§1.1中介绍数理统计学的任务时, 我们谈到了收集和使用数据这两个方面. 并说明,所谓"使用",即将样本用于解决所提出的问题,作出一定的结论,叫**统计推断**. 现在我们可对此作出更仔细而严密的解释. 本节的另一任务是, 通过一些例子说明统计推断的种种形式,借以介绍数理统计学的若干主要分支.

(一) 样本分布族,参数和参数空间

在§1.2(二)中我们反复申述并举例说明了这样的观点: 统计模型就是样本分布: 后者包含了样本中的全部信息; 所要解决的问题必须与样本分布有某种关联. 如在例1.4和例1.5中, 若知道了 a 或 λ, 则问题已解决而没有必要抽样了. 正因为 a 和 λ 未知, 才需要通过抽样去了解它. 在这两个例子中, 抽样的目的都是为了解出现在抽样分布中的某些未知常数.

在统计上, 把出现在样本分布中的未知常数称为**参数**. 如在例1.4中, 问题在于估计 a, a 当然未知, 是一个参数. 至于 σ 是否是参数,则要看情况. 若我们对这天平的精度已有足够的了解, 而可以给出 σ 的值(如 $\sigma=1$), 则 σ 不是参数. 若并无足够了解, 甚至问题就在于估计天平的精度, 则 σ 自然是参数. 这时可称 (a, σ) 为参数向量. 例1.5中的参数为 λ, 而例1.6有五个参数: a, b, σ_1, σ_2, ρ.

在一具体问题中,样本分布中的参数所取的值虽未知,但根据该参数的性质, 可以给出参数值所在的范围. 这个范围叫做**参数空间**. 如在例1.4中,若 a、σ 都是参数, 由 σ 的性质知它必大于0. 至于 a,在本问题中作为物件的重量,也只能大于0. 因此参数

空间可取为

$$\Theta=\{(a,\ \sigma):a>0,\ \sigma>0\},$$

即平面上的第一象限．根据我们事前的了解，也可能把本问题参数空间限制在一个更小的范围内，如 $\{(a,\ \sigma):0<a_1\leqslant a\leqslant a_2<\infty,\ 0<\sigma<\sigma_0\}$．参数空间的这种取法表明：在没有进行试验（用天平秤）之前，我们就有把握肯定物件重量不小于 a_1 也不大于 a_2．同样，在例 1.5 中参数空间是 $\Theta=\{\lambda:\lambda>0\}$，而在例 1.6 则是

$$\Theta=\{(a,\ b,\ \sigma_1,\ \sigma_2,\ \rho):\ a>0,\ b>0,\ \sigma_1>0,\ \sigma_2>0,\ |\rho|\leqslant 1\},$$

$|\rho|\leqslant 1$ 是根据 ρ 是身高与体重的线性相关系数，而后者必在 -1 与 1 之间．在本例中我们有把握肯定 ρ 非负，故取 $\rho>0$ 也可以．

样本分布中既然包含了一些未知的参数，那么，可能的样本分布就不止一个，而是一个**分布族**．例如，例 1.5 的样本分布族是

$$\{f(x,\ \lambda):\ \lambda>0\}.$$

要注意此式中 x 和 λ 的质的区别：λ 是未知常数，它的每一个可能值对应于一个具体的样本分布，x 只是密度函数中的一个流动变量，有如定积分中的积分变量，无甚特别含义．

现在我们可以把 §1.2（三）中所给的统计模型的定义稍稍确切化一点：统计模型就是样本分布族．这一点的实际意义在于：样本分布族，连同其参数空间，从总的方面定出了问题的范围．也可以这样说：分布族反映了因我们对所研究问题以及抽样方式的知识和规定，我们能把问题确定[1]到何种程度．分布族愈小，确定的程度愈高．这一般就意味着愈有可能作出更为精确和可靠的结论．拿例 1.4 来说，设问题在于估计重量 a．若反映天平精度的 σ 未知，则分布族大些．若 σ 已知（例如 $\sigma=1$），则分布族小些．可以想象：有关天平精度的知识，对估计 a 是有用的．这可以说成：在知道 σ 时，估计 a 的问题更确定一些．

在以上诸例中，参数都取实数值，而参数空间则是欧氏空间的

1) 在此要注意，"确定"一词不是指问题是不是提清楚了．问题提法总应是"清楚"的．这里意思是：问题中已知的成分愈多，问题就愈确定，推断未知的成分就愈容易．

一部分．这种情况下的统计问题称为**参数统计问题**．在有些问题中情况较此复杂，我们举一个例子：

例 1.8 我们再来考察例 1.4 中估计物件重量 a 的问题．若我们对秤量的随机误差的性质不甚了解，或者说，有理由怀疑，例 1.4 中所论述的、导致随机误差服从正态分布的根据可能不成立，则我们不能认定样本分布为正态的．但可能有根据作以下的一般性假定：秤量的随机误差有均值 0（天平无系统误差），随机误差的分布关于 0 对称，且有密度函数．又假定，例 1.4 中所述关于样本 $X_1, \cdots, X_n$ 独立同分布的道理仍有效，则 $(X_1, \cdots, X_n)$ 的分布有概率密度

$$f(x_1-a)f(x_2-a)\cdots f(x_n-a), \tag{1.14}$$

其中 a 为一个实参数，取值于 $(0, \infty)$．f 为一个概率密度函数，满足条件

$$f(-x)=f(x) \text{ 对一切 } x\in(-\infty, \infty), \int_{-\infty}^{\infty}|x|f(x)dx<\infty. \tag{1.15}$$

在这里，分布 (1.14) 中的未知成分有二：一是实参数 a，另一是密度 f．后者不是实参数，它变化的范围由条件 (1.15) 规定．整个来说，样本分布的参数不能由有限个实参数来刻划，参数空间也不是欧氏空间的一部分．若以 $\mathscr{F}$ 记由 (1.15) 所确定的概率密度的集，则形式上可把"参数空间"定义为

$$\Theta=\{(a, f): a>0, f\in\mathscr{F}\}. \tag{1.16}$$

这类情况（即参数空间不是欧氏空间的一部分）下的统计问题称为**非参数统计问题**．

（二）统计推断

有了以上的准备，现在可以来较确切地回答什么是统计推断问题．为解决具体问题而抽样，样本分布族规定了问题的统计模型，样本的具体数值供实际作统计推断之用．推断什么？就是推断样本分布中的未知参数．根据问题的需要，可以只对一部分未

知参数作推断,推断的形式也根据问题的需要而不同. 如在例 1.4 中,问题是要估计物重 a. 用统计的语言,就是依据样本 $X_1, \cdots, X_n$ 对分布 $N(a, \sigma^2)$ 中的参数 a 作出推断,这个推断的具体内容就是依据样本 $X_1, \cdots, X_n$ 算出一个数,以之作为 a 的估计值. 也可能问题只要求判定"物重 a 是否超过 1 公斤",则推断的内容是建立一个法则,按这个法则,对每组具体样本 $X_1, \cdots, X_n$,在"a 是超过 1 公斤"和"a 不超过 1 公斤"这两个判定中采取一个. 本问题不直接涉及 σ,故没有推断 σ 的问题.

"推断"一词在日常使用中,是指从一定的条件和假定出发,按照一定的方法或规则,而得出未知事物的某种结论. 统计推断也是这种模式:"未知事物"就是未知参数,"一定的条件和假定"就是样本和统计模型.

人们常把统计推断说成是由样本推断总体,或由部分推断整体. 这是在所掌握的知识不完全的情况下,所作的一种归纳性的推理,不同于在几何学中证明"等腰三角形底角相等"或在代数学中求一个二次方程的根. 在这些问题中,推理的步骤在有关公理体系之下是公认的,结果是确切的,统计推断则不然. 设想从一批 10,000 件产品中抽出 100 个以估计整批产品的废品率,我们掌握的知识很不完全——有 9900 件的情况全然未知,这自然就排除了通过一种严格的推理以得出正确结果的可能性. 在各种活动领域,以至在日常生活中,人们都常面临这样的情况:掌握的知识不完全,但需要作结论. 统计推断问题的特点在于:所掌握的这部分知识(样本),通过概率的方式与整体发生了一种有严格数学意义的联系. 这种联系可以使统计推断能提成一定形式的数学问题,而不能改变这个根本之点——统计推断的结果不能保证不错. 如在例 1.4 中估计物重 a. 有了样本以后,可以用种种方法,例如 §1.1 (一)3 中给出的几种方法,去估计 a. 但我们不仅不能保证这些以及其他任何方法估计的结果恰为 a,也无法保证绝对不会发生较大的偏差. 在判定 a 是否超过一公斤的问题中,不论用什么方法也无法保证不会错判.

但这样就产生一个问题：统计推断有什么意义？统计推断，以至于数理统计学，能否算是一门科学？回答是肯定的，理由在于下面两点：1．虽然不能保证在每一具体情况下所作的统计推断不错，但使用概率论的方法，可以给出种种有意义的指标，去衡量推断的正确程度．这可以计算出来．比方说，在例1.4估计 a 的问题，在所设正态模型下，若用样本的算术平均值 $\bar{x}=\sum_{i=1}^{n} X_i/n$ 去估计 a，则我们可算出：$\bar{x}$ 与正确值 a 的偏离超过给定的数 $c>0$ 的机会有多大．这个机会，即概率 $P(|\overline{X}-a|>c)$，可作为 $\bar{x}$ 这个推断的正确性的一个合理指标（c 为容许的误差限）．在此，我们进一步明确了数理统计学与概率论的密切关系．不仅统计问题的提法有赖于概率论，统计推断的方法、结果的表述形式，以及衡量一个统计推断的优劣等，都离不开概率论．2．统计推断的数学理论是一种严整的理论，符合现代数学的严格性标准．这与统计推断的结果在具体问题中可能有错是两回事．还是拿例1.4估计 a 的问题来谈．在§1.1（一）3中我们提出了三种估计方法，那一个最好？在什么意义下最好？可以用严格的数学方法证明：在种种合理的优良性准则之下，当统计模型确为正态时，样本算术平均这个估计最优．这是一个严格的数学命题．不言而喻，这个"最优"，必然是在整体即概率的意义下，对一组具体的数据而言，样本算术平均与真值 a 的偏差，完全可以大于其他两个估计的偏差．

我们上面提出的这两点也回答了这个问题：统计推断包含那些内容？归纳起来有三点：提出种种具体的统计推断方法；计算有关这些方法的性能的种种数量指标（例如上文的 $P(|\overline{X}-a|>c)$；在一定条件和一定意义下寻找最优的推断方法，或证明某种统计推断方法是最优的．这三方面当然不是截然分开，而是密切联系在一起的．

数理统计学的基础是概率论，统计推断从结论的表述到优良性的刻划，都离不开概率．但概率的本质是什么？这个带哲学性的问题在学者中的认识并不一致，而这并非与应用无关．一种流行

的、在直观上容易理解的说法，是通过频率去解释概率．按照这种解释，如果说某一统计推断方法在某种情况下的正确性为 90%，其意义是：当将这一推断方法在同样情况下反复使用大量次数时，平均每 100 次中有 90 次能给出正确的推断．倘若问题的性质不允许有什么重复，则这个解释就失去了意义．以这种观点去处理和解释统计推断，就叫做频率学派，也有称为古典学派的．这一学派在数理统计发展的较早时期(比方说，本世纪五十年代之前)占据了统治地位，目前也仍占优势．但其他的学派，主要是 Bayes 学派(见第五章)，也正以很大势头发展起来，也还有若干影响较小，但也值得注意的观点，其中有些将在本书相应的地方提到．

§1.4 统计量和抽样分布

(一) 统计量

前几节中我们反复申述了这样的观点：为了研究一个问题而收集数据，数据就是样本．通过确定样本分布而建立统计模型．但要具体地实施统计推断，则要依据表现为具体数值的样本．样本自身是一堆杂乱无章的数字，要对这些数字进行加工整理，计算出一些量以用于统计推断．可以这样理解：这种由样本计算出来的量，把样本中与所要解决的问题有关的信息集中起来了．如在例 1.4 中，为估计物重 a，我们可以使用样本的算术平均值 $\bar{X}$，或者其他某个看来合理的量，如在 §1.1 (一) 3 中所指出的几种，它们都是由样本算出的．

在统计上，把凡是由样本算出的量称为**统计量**，或者说：统计量就是样本的函数．在此十分重要的一点是：统计量只依赖于样本，而不能与任何未知的量有关．特别是：统计量不能依赖于未知参数，这一点从统计量的意义看应该是显然的，因为统计量的作用就在于对未知参数进行推断．在什么问题中选用什么统计量，当然要看问题的性质．笼统地可以说：所提出的统计量应是最好地集中了与问题有关的信息．这不见得总是容易做到的．往往最初

是从直观或某些一般性原则考虑提出统计量，再考察它是否在某种意义下较好地集中了样本中与问题有关的信息量.

下面举几个常用的重要统计量的例子：

例 1.9 样本均值 设 $X_1, \cdots, X_n$ 为样本，则

$$\overline{X}=\sum_{i=1}^{n} X_i/n \tag{1.17}$$

称为**样本均值**. 在样本 $X_1, \cdots, X_n$ 为独立同分布的情况下，样本均值常用于估计总体分布的均值，或检验有关总体分布均值的假设.

例 1.10 样本方差 设 $X_1, \cdots, X_n$ 为样本，则

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(\overline{X}-X_i)^2 \tag{1.18}$$

称为**样本方差**. 在样本 $X_1, \cdots, X_n$ 为独立同分布的情况下，样本方差可用于估计总体分布的方差. (1.18)式中的 $n-1$ 称为 S^2 的**自由度**. 自由度这个名词的解释有三：1. 一共有 n 个数值 $X_1, \cdots, X_n$ 应该有 n 个自由度(因每个样本都可自由变化，不受其他样本的牵连)，但有一个自由度已用于估计总体分布均值(用 $\overline{X}$)，还剩 $n-1$ 个自由度. 2. S^2 是 n 个数 $X_1-\overline{X}, \cdots, X_n-\overline{X}$ 的平方和，但这 n 个数受到一个(也只一个)约束，即 $\sum_{i=1}^{n}(X_i-\overline{X})=0$，故只有 $n-1$ 个自由度. 3. 若以 $\overline{X}=\sum_{i=1}^{n}X_i/n$ 代入 $\sum_{i=1}^{n}(X_i-\overline{X})^2$，而将其整理为二次型 $\sum_{i,j=1}^{n}a_{ij}X_iX_j(a_{ij}=a_{ji})$，则不难验证：方阵 $A=(a_{ij})$ 的秩为 $n-1$. 自由度就定义为这个秩.

自由度是数理统计中常见的名词. 以上三种解释，前两种较形象化，而最后一种在数学上最清楚严谨，适用于一切与样本二次型有关的统计量.

例 1.11 次序统计量及与之有关的统计量 设 $X_1, \cdots, X_n$ 为样本，把 $X_1, \cdots, X_n$ 按由小到大的次序排列成 $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$，则 $(X_{(1)}, \cdots, X_{(n)})$ 称为**次序统计量**. 例如，若 $X_1=$

1.3, $X_2=0.7$, $X_3=1.8$, 则次序统计量为(0.7, 1.3, 1.8). 单个的 $X_{(i)}$, 或($X_{(1)}, \cdots, X_{(n)}$)的一部分, 也称为次序统计量, 这不致引起混淆.

通过次序统计量可定义一些在实用上有重要意义的统计量:

1. 样本中位数 是由下式定义的 m:

$$m=\begin{cases}X_{\left(\frac{n+1}{2}\right)}, & \text{当 } n \text{ 为奇数;}\\ \dfrac{1}{2}\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right), & \text{当 } n \text{ 为偶数,}\end{cases} \tag{1.19}$$

就是次序统计量中位置在正中的那一个, 或位置最靠中的那两个的平均. 在 $X_1, \cdots, X_n$ 为独立同分布样本时, m 可用于估计总体分布中位数. 若已知总体分布关于某点对称, 则对称中心既是中位数也是均值, 于是 m 也可作为总体分布均值的估计. 正态总体就是一个例子.

2. 样本 p 分位数($0<p<1$) 可定义为 $X_{[(n+1)p]}$, 此处$[a]$表示不超过 a 的最大整数. 当$p=\dfrac{1}{2}$而 n 为奇数时, 此定义与(1.19)一致, 但 n 为偶数则不然. 可把样本 p 分位数的定义加以适当修改, 以使当 $p=1/2$ 时与(1.19)一致. 当 n 较大时这个修改是很有限的, 一般无此必要. 在 $X_1, \cdots, X_n$ 为独立同分布样本时, 样本 p 分位数可用于估计总体分布的 p 分位数.

3. 极值 指 $X_{(n)}$ 和 $X_{(1)}$, 分别称为样本极大值和样本极小值. 样本极值在某些关于灾害性现象和材料试验结果的统计分析中有用. 如一定时期内一条河的最大流量、最大地震震级、材料断裂强度等, 都是极值性的量. 在数理统计中有一个叫**极值统计分析**的专题处理这种问题, 后者也可视为次序统计量的统计分析的一部分. 在极值统计分析中也用到较次要的极值, 即 $X_{(2)}, X_{(3)}, \cdots, X_{(n-1)}, X_{(n-2)}, \cdots$ 等.

4. 极差 即 $X_{(n)}-X_{(1)}$. 在样本 $X_1, \cdots, X_n$ 独立同分布时, 极差可用于估计总体分布的散布程度.

其他重要的统计量将在本书适当的地方引进来.

（二）抽样分布

样本是随机变量，有一定的概率分布，即样本分布．统计量既是样本的已知函数，则它也有其概率分布，且这个概率分布在原则上可由样本分布定出．统计量的概率分布称为（该统计量的）**抽样分布**．

确定种种统计量的抽样分布，是数理统计学的一个基本问题．在§1.1（三）中提到的R. A. Fisher 1922年那篇著名论文«On he mathematical foundations of stratistics»中，Fisher把数理统计学的任务概括为三条：1．"specification"，即定模型，确定样本分布（如前指出的，在独立同分布样本的场合，也可理解为总体分布）．2．"estimation"，即估计，用样本估计模型中的未知参数．3．"sampling distribution"，即抽样分布．这问题的重要性是很明显的：统计推断的结果取决于抽得的样本，而样本受到随机性的干扰，因而推断的结果也是随机的：一个整体上看来较好的推断方法，在个别情况下可以给出不好的结果．反之亦然．因此，统计推断方法优良性的指标只能是整体性的，即取决于所用统计量的抽样分布．总之，要想得到一种特定的统计推断方法的性能的全面了解，必须确定其抽样分布．如在例1.4中，用样本均值$\overline{X}$估计正态分布的均值a．$\bar{x}$与a的偏差超过一定限度的机会多大，即概率$P(|\overline{X}-a|>c)$．要算出这个概率，需要知道$\overline{X}$的抽样分布．

使用概率论中的已知结果，不难在某些简单的统计模型中，定出某些简单的统计量的抽样分布．举几个例子．

例1.12 设$X_1, \cdots, X_n \sim N(a, \sigma^2)$（记号意义见(1.10)），求$\overline{X}$的抽样分布．一般可考虑$T=\sum_{i=1}^{n} a_i X_i$，$a_i$都是常数，$\overline{X}$相当于$a_1=\cdots=a_n=\frac{1}{n}$的特例．根据$X_1, \cdots, X_n$独立同分布且$X_1 \sim N(a, \sigma^2)$，由概率论中熟知的结果，即知$T$的分布为$N(a\sum_{i=1}^{n} a_i, \sigma^2\sum_{i=1}^{n} a_i^2)$，故$\overline{X}$有分布$N(a, \sigma^2/n)$，记为$\overline{X}\sim N(a, \sigma^2/n)$．

$\bar{X}$ 的分布与参数 a, σ 有关(某些统计量也可以只依赖于一部分参数. 例如，统计量 X_1-X_2 的分布为 $N(0, 2\sigma^2)$，与 a 无关). 因此有的读者可能对这一点有疑问: 我们强调统计量只依赖于样本而与参数无关，为什么其分布又与参数有关? 回答是明显的: 统计量的表达式固然只与样本有关，但样本的分布依赖于参数. 不仅如此，我们还得强调: *有用的统计量的抽样分布必须与参数有关*. 不然的话，该统计量就不包含有关参数的任何信息，对推断这个参数毫无用处. 如上文，统计量 X_1-X_2 的分布与 a 无关. 但它无助于估计 a，这在直观上看很明显.

例 1.13 设 $X_1, \cdots, X_n \sim f(x, \lambda)$，$f(x, \lambda)$ 由(1.8)式定义，求 $\bar{X}$ 的样本分布.

以 $f_n(x, \lambda)$ 记 $T_n=X_1+\cdots+X_n$ 的密度函数(注意 $f_1(x, \lambda)$ 即为 $f(x, \lambda)$). 有 $T_n=T_{n-1}+X_n$. 用独立和的密度公式有

$$f_n(x, \lambda)=\int_{-\infty}^{\infty} f(x-y, \lambda) f_{n-1}(y, \lambda) dy$$

$$=\int_{-\infty}^{x} \lambda e^{-\lambda(x-y)} f_{n-1}(y, \lambda) dy.$$

以此，通过简单的积分计算，不难用归纳法证明

$$f_n(x, \lambda)=\begin{cases} \dfrac{1}{(n-1)!} \lambda^n e^{-\lambda x} x^{n-1}, & \text{当 } x>0; \\ 0, & \text{当 } x \leqslant 0. \end{cases} \tag{1.20}$$

因而得出 $\bar{X}=T_n/n$ 的概率密度 $g_n(x, \lambda)$ 为

$$g_n(x, \lambda)=n f_n(nx, \lambda)=\begin{cases} \dfrac{1}{(n-1)!} n^n \lambda^n e^{-\lambda n x} x^{n-1}, & \text{当 } x>0; \\ 0, & \text{当 } x \leqslant 0. \end{cases} \tag{1.21}$$

例 1.14 一批产品有 N 个，内废品 M 个，从其中抽出 n 个，抽样方式是: 1. 随机，无放回; 2. 随机，有放回. 以 T 记抽出的这 n 个中废品的个数，T 是一统计量，求其抽样分布.

这问题在初等概率论中已解决了，结果是:

情况 1　超几何分布:

$$P(T=x)=\binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n} \tag{1.22}$$

$\left(\text{当 } b<0 \text{ 或 } b>a \text{ 时, 约定}\binom{a}{b}=0\right)$.

情况 2　二项分布 $B(n, p)$, $p=M/N$:

$$P(T=x)=\binom{n}{x}p^x(1-p)^{n-x},\ x=0,\ 1,\ \cdots,\ n. \tag{1.23}$$

例 1.15　设 $X_1, \cdots, X_n \sim F$, $1\leqslant i\leqslant n$, 求次序统计量 $X_{(m)}$ 的抽样分布.

为要事件 $\{X_{(m)}<x\}$ 发生, 必须只须 n 个事件 $\{X_i<x\}$, $i=1, \cdots, n$, 至少发生 m 个. 这 n 个事件相互独立且都有概率 $F(x)$, 故 $P(X_{(m)}<x)=\sum_{i=m}^{n}\binom{n}{i}F^i(x)[1-F(x)]^{n-i}$. 不难证明(习题8), 此式可改写为积分形式

$$P(X_{(m)}<x)=m\binom{n}{m}\int_0^{F(x)}t^{m-1}(1-t)^{n-m}dt. \tag{1.24}$$

若 F 有密度 f, 则 $X_{(m)}$ 也有密度, 且等于

$$f_m(x)=n\binom{n}{m}F^{m-1}(x)[1-F(x)]^{n-m}f(x). \tag{1.25}$$

若 F 有密度 f, 则次序统计量的联合密度为

$$g(x_{(1)},\ \cdots,\ x_{(n)})=\begin{cases} n!f(x_{(1)})\cdots f(x_{(n)}), & \text{当 } x_{(1)}<\cdots<x_{(n)}; \\ 0 & \text{, 其他情况,} \end{cases}$$

由此可得出任意个次序统计量的联合密度. 比方说, 固定 $x_{(m)}=x$, 而对 $x_{(1)}, \cdots, x_{(m-1)}$ 在 $x_{(1)}<\cdots<x_{(m-1)}<x$ 内积分, 对 $x_{(m+1)}, \cdots, x_{(n)}$ 在 $x<x_{(m+1)}<\cdots<x_{(n)}$ 内积分, 不难得到 $X_{(m)}$ 的密度 $f_m(x)$ 如 (1.25). 又如, 固定 $x_{(i)}=x$, $x_{(j)}=y$ $(i<j,\ x<y)$, 对 $x_{(1)}, \cdots, x_{(i-1)}, x_{(i+1)}, \cdots, x_{(j-1)}, x_{(j+1)}, \cdots, x_{(n)}$ 在 $x_{(1)}<\cdots<x_{(i-1)}<x$, $x<x_{(i+1)}<\cdots<x_{(j-1)}<y$, $y<x_{(j+1)}<\cdots<x_{(n)}$ 内积分, 不难算出 $(X_{(i)}, X_{(j)})$ 的联合密度为

$$f_{ij}(x,\ y)=\begin{cases}\dfrac{n!}{(i-1)!(j-i-1)!(n-j)!}F^{i-1}(x)[F(y)\\ \qquad -F(x)]^{j-i-1}[1-F(y)]^{n-j}, & \text{当 } x<y;\\ 0, & \text{当 } x\geqslant y.\end{cases}\tag{1.26}$$

令 $V=X_{(j)}-X_{(i)}$，由(1.26)不难推出 V 的密度．事实上，作变换

$$V=X_{(j)}-X_{(i)},\ Z=X_{(i)}.$$

逆变换为 $X_{(i)}=Z,\ X_{(j)}=V+Z$．变换的 Jacobi 为 1．由(1.26)得$(V,\ Z)$的密度为

$$g(v,\ z)=\begin{cases}\dfrac{n!}{(i-1)!(j-i-1)!(n-j)!}F^{i-1}(z)[F(v+z)\\ \qquad -F(z)]^{j-i-1}[1-F(v+z)]^{n-j}, & \text{当 } v>0;\\ 0, & \text{当 } v\leqslant 0.\end{cases}$$

由此得出 V，即 $X_{(j)}-X_{(i)}$，有密度 $\int_{-\infty}^{\infty}g(v,\ z)dz$．特别，若 F 是 $R(0,\ 1)$，即$(0,\ 1)$内均匀分布，则 $F(x)=0,\ x$ 或 1，视 $x\leqslant 0,\ 0<x<1$ 或 $x>1$ 而定．对这个 F，上述积分不难算出，结果为

$$g_{nij}(v)=\begin{cases}\dfrac{n!}{(j-i-1)!(n-j+i)!}v^{j-i-1}(1-v)^{n-j+i}, & 0<v<1;\\ 0, & \text{其他 } v.\end{cases}\tag{1.27}$$

(三) χ^2、t 和 F 分布

能算出抽样分布的确切而简单的表达式的情况，为数不多．所幸的是，有一个重要情况，即总体分布为正态时，许多重要统计量的抽样分布已经得出．这些多与本段标题中指明的这三种分布有密切关系．

1．χ^2 分布　设 $X_1,\ \cdots,\ X_n$ 为独立同分布的随机变量，且

$X_1 \sim N(0, 1)$. 令 $\xi=\sum_1^n X_i^2$. 则 ξ 的分布称为具自由度 n 的 χ^2 分布（注意二次型 $\sum_{i=1}^n X_i^2$ 的矩阵为 n 阶单位阵，其秩为 n. 因此，这里自由度的意义就是如例 1.10 中所说的)，记为 $\xi \sim \chi_n^2$.

为求 ξ 的概率密度 $f(x)$，先计算 $P(\xi<x)$. 当 $x \leqslant 0$ 时此概率显然为 0. 当 $x>0$ 时，依定义有

$$P(\xi<x)=(2\pi)^{-n/2}\int\cdots\int_B \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right) dx_1 \cdots dx_n.$$

此处 B 为球体 $\{(x_1, \cdots, x_n): \sum_{i=1}^n x_i^2<x\}$. 转换到球坐标，有

$$P(\xi<x)=c_n \int_0^{\sqrt{x}} e^{-r^2/2} r^{r_1-1}\, dr,$$

此处 c_n 为某个与 n 有关的常数. 为求出 c_n，利用 $\lim_{x\to\infty} P(\xi<x)=1$ 的事实，由上式得

$$1=c_n \int_0^{\infty} e^{-r^2/2} r^{n-1}\, dr.$$

在此式中作变数代换 $r=\sqrt{2t}$，再利用 Gamma 积分，易算出 $c_n=2^{1-n/2}/\Gamma\left(\frac{n}{2}\right)$. 以此代入 $P(\xi<x)$ 表达式，对 x 求导，即得 ξ 的概率密度为

$$f(x)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{\frac{n}{2}-1}, & \text{当 } x>0; \\ 0, & \text{当 } x \leqslant 0.\end{cases} \tag{1.28}$$

它称为自由度 n 的 χ^2（分布）密度.

下面要证明 χ^2 分布的几条性质，它们在统计上有重要应用. 这依赖于下面的预备事实：

引理 1.1 设 $X_1, \cdots, X_n$ 相互独立，X_i 服从正态分布 $N(a_i, \sigma^2)$，$i=1, \cdots, n$. 设 $A=(a_{ij})$ 为 n 阶正交方阵，$Y_i=\sum_{j=1}^n a_{ij}X_j$，$i=1, \cdots, n$. 则 $Y_1, \cdots, Y_n$ 独立，$Y_i \sim N(b_i, \sigma^2)$，$b_i=\sum_{j=1}^n a_{ij}a_j$，$i=1, \cdots, n$.

证 由引理假定，知 $(X_1, \cdots, X_n)$ 的联合密度为 $(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a_i)^2\right)$. 因$A$为正交阵,变换的 Jacobi 行列式绝对值为1. 又由 A 的正交性有 $\sum_{i=1}^{n}(y_i-b_i)^2=\sum_{i=1}^{n}(x_i-a_i)^2$. 于是由多维密度变换公式得 $(Y_1, \cdots, Y_n)$ 有概率密度 $(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-b_i)^2\right)$. 这就证明了引理的全部论断.

现往下证明

定理 1.1 若 $X_1, \cdots, X_n \sim N(a, \sigma^2)$, $\bar{X}$ 和 S^2 分别是样本均值和样本方差(见例 1.9, 1.10). 则1. $\bar{X}\sim N(a, \sigma^2/n)$; 2. $(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$; 3. $\bar{X}$ 与 S^2 独立.

证 取正交阵 A (n 阶), 使其第一行 $(a_{11}, \cdots, a_{1n})$ 为 $\left(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}}\right)$. 作变换 $Y=AX$, $Y=(Y_1, \cdots, Y_n)'$, $X=(X_1, \cdots, X_n)'$[1]. 则依引理1.1, $Y_1, \cdots, Y_n$ 独立, $Y_i\sim N(b_i, \sigma^2)$, $b_i=a\sum_{j=1}^{n}a_{ij}$. 有 $b_1=\sqrt{n}\,a$. 由 A 为正交阵知 $\sum_{j=1}^{n}a_{ij}=0$ 当 $i\geqslant 2$, 故 $b_i=0$ 当 $i\geqslant 2$. 又由正交性有

$$(n-1)S^2=\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}X_i^2-n\bar{X}^2$$

$$=\sum_{i=1}^{n}Y_i^2-Y_1^2=\sum_{i=2}^{n}Y_i^2.$$

这里已用了 $\bar{X}=Y_1/\sqrt{n}$. 因而 $\bar{X}\sim N(a, \sigma^2/n)$(这已在例1.12中证过). 因为 $Y_2/\sigma, \cdots, Y_n/\sigma$ 为服从 $N(0, 1)$的独立同分布变量,由 χ^2 分布的定义知$(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$. 最后,由于 $\bar{X}$ 只依赖 Y_1, S^2 只依赖$(Y_2, \cdots, Y_n)$,而 $Y_1, \cdots, Y_n$ 全体独立,知 $\bar{X}$ 与 S 独立. 定理证毕.

本定理对正态分布参数的统计推断有重大意义.

定理 1.2 (Cochran 定理) 设 $X=(X_1, \cdots, X_n)'$, $X_1, \cdots,$

---

1) 本书中以不加“′”的向量为列向量,加“′”表示转置.

X_n 相互独立，$X_i \sim N(a_i, \sigma^2)$，$i=1, \cdots, n$. 又设 $A_1, \cdots, A_m$ 都是 n 阶非负定方阵，$A_1+\cdots+A_m=I_n$(I_n 为 n 阶单位阵). 令 $\xi_i=X'A_iX$，$i=1, \cdots, m$，则当

$$\sum_{i=1}^{m} rk(A_i)=n \quad (rk(A_i) \text{为} A_i \text{的秩}) \tag{1.29}$$

时，$\xi_1, \cdots, \xi_m$ 相互独立. 若记 $a=(a_1, \cdots, a_n)$，则当(1.29)成立且 $a'A_1a=0$ 时，$\xi_1/\sigma^2 \sim \chi^2_{n_1}$，$n_1=rk(A_1)$.

证 记 $n_i=rk(A_i)$. 因 A_i 为秩 n_i 的非负定 n 阶方阵，由矩阵论知存在 $n\times n_i$ 矩阵 B_i，使 $A_i=B_iB_i'$. 记 $B=(B_1 \vdots B_2 \vdots \cdots \vdots B_m)$，则 B 为 n 阶方阵. 作变换 $Y=(Y_1, \cdots, Y_n)'=B'X$，则因 $\sum_{i=1}^{m} A_i=I_n$，有

$$Y'Y=X'BB'X=X'\sum_{i=1}^{m} B_iB_i'X=\sum_{i=1}^{m} X'A_iX=X'X.$$

由此知 B 为正交阵. 故按定理1.1，知 $Y_1, \cdots, Y_n$ 独立，$Y_i \sim N(b_i, \sigma^2)$，$i=1, \cdots, n(b=(b_1, \cdots, b_n)=B'a)$. 注意到

$$B_i'X=(Y_{n_1+\cdots+n_{i-1}+1}, \cdots, Y_{n_1+\cdots+n_i}),\ i=1, \cdots, m(n_0=0),$$

知 $\xi_i=\sum_{n_1+\cdots+n_{i-1}+1}^{n_1+\cdots+n_i} Y_j^2$，$i=1, \cdots, m$. 于是由 $Y_1, \cdots, Y_n$ 相互独立，推出 $\xi_1, \cdots, \xi_m$ 相互独立.

又若 $a'A_1a=0$，则 $a'B_1B_1'a=0$，因而 $B_1'a=0$，即 $b_1=\cdots=b_{n_1}=0$. 于是 $\xi_1/\sigma^2=\sum_{j=1}^{n_1}(Y_j/\sigma)^2$，其中 $Y_1/\sigma, \cdots, Y_{n_1}/\sigma$ 相互独立，各有正态分布 $N(0, 1)$. 这证明了 $\xi_1/\sigma^2 \sim \chi^2_{n_1}$. 定理证毕.

此定理在本书第六、七章中有重要应用. 作为推论，有下面的结果：

系1.1 设 $\xi_1, \cdots, \xi_m$ 相互独立，$\xi_i \sim \chi^2_{n_i}$，$i=1, \cdots, m$，则 $\xi=\sum_{i=1}^{m} \xi_i \sim \chi^2_{n_1+\cdots+n_m}$.

这当然也不难直接由定义证明.

2. t 分布　设随机变量 X、Y 独立，$X \sim N(0, 1)$，$Y \sim \chi^2_n$. 令 $\xi=X\Big/\sqrt{\frac{1}{n}Y}$. 则 ξ 的分布称为**具自由度 n 的 t 分布**（这里，ξ

的自由度取为 Y 的自由度)，记为 $\xi\sim t_n$.

由 Y 的密度为(1.28)，易得出 $\sqrt{\frac{1}{n}Y}$ 的密度为 $h_2(x)=2n^{n/2}x^{n-1}e^{-nx^2/2}\Big/\left(2^{n/2}\Gamma\left(\frac{n}{2}\right)\right)$(当 $x>0$). 于是 ξ 的分子分母独立，且密度分别为 $h_1(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ 和 $h_2(x)$. 依概率论中熟知的关于商的密度的公式

$$g(x)=\int_0^\infty th_1(xt)h_2(t)dt, \tag{1.30}$$

易算出 ξ 的概率密度函数为

$$g(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\ -\infty<x<\infty. \tag{1.31}$$

这个密度函数的图形与标准正态分布密度函数 $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ 的图形相似：都是关于0对称，并当 $|x|\to\infty$ 时单调下降地趋于0. 在数学上可以严格证明：当自由度 $n\to\infty$ 时，自由度为 n 的 t 分布收敛于标准正态分布 $N(0, 1)$.

3. **F 分布** 设随机变量 X、Y 独立，$X\sim\chi_m^2$，$Y\sim\chi_n^2$. 令 $\xi=\frac{1}{m}X\Big/\frac{1}{n}Y$. 则 ξ 的分布称为具**自由度 m 和 n 的 F 分布**(注意分子的自由度在前)，记为 $\xi\sim F_{m,n}$.

由此定义立见：若 $\xi\sim t_n$，则 $\xi^2\sim F_{1,n}$.

$F_{m,n}$ 的概率密度函数不难算出. 因为由(1.28)知 X, Y 的密度，故 X/m 和 Y/n 的密度容易算出. 再由分子分母独立，按公式(1.30)，即算出密度函数为

$$h(x)=\begin{cases}\dfrac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(m/2)\Gamma(n/2)}m^{m/2}n^{n/2}x^{\frac{m}{2}-1}(n+mx)^{-\frac{1}{2}(m+n)}, & x>0;\\ 0, & x\leqslant 0.\end{cases} \tag{1.32}$$

4. 非中心的 χ^2、t 和 F 分布　设 $X_1, \cdots, X_n$ 独立，$X_i \sim N(a_i, 1)$, $i=1, \cdots, n$. 令 $\xi=\sum_1^n X_i^2$, $\delta=(\sum_1^n a_i^2)^{1/2}$. 则 ξ 的分布称为具自由度 n 和**非中心参数** δ 的**非中心 χ^2 分布**，记为

$$\xi \sim \chi_{n,\delta}^2.$$

设 X, Y 独立，$X \sim N(\delta, 1)$, $Y \sim \chi_n^2$. 令 $\xi = X \Big/ \sqrt{\frac{1}{n} Y}$. 则 ξ 的分布称为具自由度 n 和非中心参数 δ 的**非中心 t 分布**，记为

$$\xi \sim t_{n,\delta}.$$

设 X, Y 独立，$X \sim \chi_{m,\delta}^2$, $Y \sim \chi_n^2$. 令 $\xi = \frac{1}{m} X \Big/ \frac{1}{n} Y$. 则 ξ 的分布称为具自由度 m, n 和非中心参数 δ 的**非中心 F 分布**，记为

$$\xi \sim F_{m,n,\delta}.$$

可以看出，前面讲过的 χ^2、t 和 F 分布，是此处当非中心参数 $\delta=0$ 时的特例，因此称为中心的 χ^2、t 和 F 分布. 在只提到这些分布的名称而未指明系中心或非中心时，一般了解为中心的. 非中心 χ^2、t 和 F 分布在统计上也有重要意义，不过就本书涉及的范围内基本上用不着它们，故不作仔细讨论了. 有些简单性质留作习题.

（四）统计量的极限分布. 大样本与小样本

当样本大小趋于无穷时，若统计量的分布趋于一定的分布，则后者称为该统计量的**极限分布**或**渐近分布**，也常称为**大样本分布**. 这可以理解为：当样本大小很大时，统计量的近似分布.

统计量的极限分布，或者更广一些，有关当样本大小趋于无穷时统计量的极限性质的研究，其意义有两个方面. 首先，如前面指出的，要弄清楚一统计推断方法的优良性如何，甚至单纯为了实施这个推断方法，往往有必要知道统计量的分布. 但后者一般很难求出，建立其极限分布，就提供了一种近似解法的可能性. 其次，统计推断方法的某些优良性准则，本身就是建立在样本大小趋于

无穷的基础上.

当样本大小趋于无穷时，一个统计量或者统计推断方法的性质，称为**大样本性质**. 大样本性质只有在样本大小趋于无穷时才有意义. 与此相对，统计量或者统计推断方法的某一性质，如果在样本大小固定时有意义，就称为**小样本性质**. 在此要强调的是，大样本和小样本的差别不在于样本个数的多少，而在于：问题是在样本大小 $n\to\infty$ 时去讨论，还是 n 固定时去讨论. 在下面将举例说明这一点. 关于大样本性质的研究构成数理统计学的一个很重要的部分，叫**大样本统计理论**. 近几十年来得到很大的发展，成为战后数理统计发展的特点之一. 有些统计分支，例如非参数统计，其中大样本理论占据了主导的地位.

大样本理论将在本书多处地方涉及，这里只举两个关于极限分布的例子.

例 1.16 设 X_1、…、$X_n \sim F$，F 有均值 a_F 和方差 σ_F^2，设 $0<\sigma_F^2<\infty$. 以 $\overline{X}_n$记样本均值 $\sum_{i=1}^{n} X_i/n$. 按 Lindeberg 中心极限定理，有

$$\sqrt{n}\,(\overline{X}_n-a_F)/\sigma_F \xrightarrow{\mathscr{L}} N(0,\ 1). \qquad (1.33)$$

这里 $\xrightarrow{\mathscr{L}}$ 表示依分布收敛. (1.33)刻划了统计量 $\overline{X}_n$ 的极限分布的形态. 这里要注意的是：$N(0,\ 1)$并非 $\overline{X}_n$ 的分布的极限，而是 $\overline{X}_n$ 经过“规则化”后的极限分布. 规则化所用常数$(a_F,\ \sigma_F)$也可以是未知的. 通常对“统计量的极限分布”一语都是作如此的理解. 事实上，由于统计量的分布依赖于未知参数，规则化所用常数自然会与这些参数有关系.

(1.33)刻划了 $\overline{X}_n$ 的一个大样本性质——**渐近正态性**. 它有其实际意义. 例如，前面曾提到：若以 $\overline{X}_n$ 作为总体均值 a_F 的估计，则 $\overline{X}_n$ 与 a_F 的偏差超过某数 c 的概率，即 $P(|\overline{X}_n-a_F|>c)$，可作为衡量这估计的优良性的一项指标. 若总体分布 F 为正态或其他简单分布，这概率值可确切地算出来. 但若 F 较复杂，甚至根本不知道其类型(非参数情况)时，这概率无法计算. 有了

(1.33)，至少在样本大小 n 较大时，可用正态分布得出这概率的近似值.

作为 a_F 的估计，$\overline{X}_n$ 还有其他一些性质. 例如，依 Kolmogorov 强大数律，有

$$P(\lim_{n\to\infty}\overline{X}_n=a_F)=1\ (\text{或写为 } \overline{X}_n\to a_F,\ \text{a.s.}) \qquad (1.34)$$

即可以用概率 1 断言：当样本大小 $n\to\infty$ 时，估计值 $\overline{X}_n$ 将任意地接近被估计值 a_F，这个性质称为 $\overline{X}_n$ 的**强相合性**. 它是一个大样本性质，因为只有在 $n\to\infty$ 时这个性质才有意义.

另一方面，我们有

$$E\overline{X}_n=a_F. \qquad (1.35)$$

就是说：估计量 $\overline{X}_n$ 的期望值，等于要估计的未知量 a_F. 这个性质称为 $\overline{X}_n$ 的**无偏性**（$\overline{X}_n$ 是 a_F 的无偏估计）. 这是一个小样本性质，因为这个性质的意义，即(1.35)，是在样本大小 n 固定时去理解的.

例 1.17 设 $X_1, \cdots, X_n\sim f$，f 为总体分布密度. 以 $X_{(m)}$ 记其次序统计量（例 1.15）. 设当样本大小 $n\to\infty$ 时，m 随 n 以一定方式变化，要求 $X_{(m)}$ 的极限分布. 我们来证明下面的定理.

定理 1.3 设 $0<p<1$，$\lim_{n\to\infty}(m-np)/\sqrt{n}=0$，$\xi_p$ 满足 $\int_{-\infty}^{\xi_p}f(x)dx=p$，且假定 f 在 ξ_p 点连续非 0（由此知 ξ_p 为 f 的唯一的 p 分位数），则当 $n\to\infty$ 时，

$$\sqrt{n}\ f(\xi_p)(X_{(m)}-\xi_p)\,/\sqrt{p(1-p)}\xrightarrow{\mathscr{L}}N(0,\ 1). \qquad (1.36)$$

证 我们前已求出 $X_{(m)}$ 的密度函数 $f_m(x)$（见(1.25)式），由此得到(1.36)式左端的密度函数 $g_m(x)$. 不难验证，$g_m(x)$ 可表为 $g_m(x)=A_1A_2$ 的形式，其中

$$A_1=\sqrt{pq}\ \sqrt{n}\binom{n}{m}F^m(t)[1-F(t)]^{n-m}\frac{m}{nF(t)},$$

$$A_2=f(t)/f(\xi_p).$$

其中 $q=1-p$, $t=\xi_p+\sqrt{\frac{pq}{n}}\frac{x}{f(\xi_p)}$. 由 f 在 ξ_p 点连续且 $f(\xi_p)>0$, 知 $\lim\limits_{n\to\infty} f(t)/f(\xi_p)=1$. 又因 $\binom{n}{m}F^m(t)[1-F(t)]^{n-m}$ 为二项分布概率, 由概率论中的局部极限定理(见复旦编《概率论》p.214),知若能证明

$$\lim_{n\to\infty}(m-nF(t))/\sqrt{npq}=-x, \qquad (1.37)$$

则有 $\lim\limits_{n\to\infty} A_1=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. 这将证明 $\lim\limits_{n\to\infty} g_m(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, 因而证明了定理 1.3. 为证(1.37) 式,只须把F在 ξ_p 点处作 Taylor 展开,得(注意 $F(\xi_p)=p$)

$$F(t)=p+\sqrt{\frac{pq}{n}}x+o\left(\frac{1}{\sqrt{n}}\right).$$

$o\left(\frac{1}{\sqrt{n}}\right)$的意义是: $\lim\limits_{n\to\infty}\sqrt{n}\,o\left(\frac{1}{\sqrt{n}}\right)=0$. 有

$$\frac{m-nF(t)}{\sqrt{npq}}=\frac{m-np}{\sqrt{npq}}-x-\frac{1}{\sqrt{pq}}\sqrt{n}\,o\left(\frac{1}{\sqrt{n}}\right).$$

于是由假定 $\lim\limits_{n\to\infty}(m-np)/\sqrt{n}=0$ 得 (1.37). 定理证毕.

不难看出: 当 $m=[(n+1)p]$ 时,定理中关于 m 的条件适合. 因此,若在(1.36)式中将 $X_{(m)}$ 改为样本 p 分位数(见例 1.11), 则(1.36)仍成立.

把 $X_1, \cdots, X_n$ 的样本中位数记为 $m_{1/2}$((1.19)式). 由定理 1.3 不难证明(习题 4)

$$2\sqrt{n}f(\xi_{1/2})(m_{1/2}-\xi_{1/2})\xrightarrow{\mathscr{L}}N(0,1). \qquad (1.38)$$

定理 1.3 的证法也可用于讨论几个次序统计量的联合极限分布的问题. 我们在此不涉及细节,只提出下面的结果:

定理 1.4 设 $0<p'<p<1$, $X_1, \cdots, X_n\sim f$, 以 ξ_p 和 $\xi_{p'}$ 记 f 的 p 和 p' 分位数,设 $f(\xi_p)>0$, $f(\xi_{p'})>0$,且 f 在 ξ_p、$\xi_{p'}$ 这两点处都连续. 以 m_p 和 $m_{p'}$ 记 $X_1, \cdots, X_n$ 的样本 p 和 p' 分位数, 则当 $n\to\infty$ 时有

$$\sqrt{n}\ (m_{p'}-\xi_{p'},\ m_p-\xi_p)' \xrightarrow{\mathscr{L}} N\left(0,\ 0,\ \frac{p'(1-p')}{f^2(\xi_{p'})},\right.$$

$$\left.\frac{p(1-p)}{f^2(\xi_p)},\ \sqrt{\frac{p'(1-p)}{(1-p')p}}\right). \tag{1.39}$$

此处 $N(0,\ 0,\ \sigma_1^2,\ \sigma_2^2,\ \rho)$ 表示一个二维正态分布，其两分量的均值都为 0，方差分别为 σ_1^2 和 σ_2^2，而两分量的相关系数为 ρ。

（五）充分统计量

统计量的充分性是数理统计学的一个重要的基本概念．它是 R. A. Fisher 在其 1922 年奠基性工作（§ 1.4（二）中所引的）中正式提出来的．但这个重要概念，Fisher 早在 1920 年以前的几年里就开始形成了．当时他与天文学家 Eddington 争论这样一个问题：设 $X_1,\ \cdots,\ X_n\sim N(a,\ \sigma^2)$，要估计反映测量精度的 σ．当时较常用的估计有两个：一是样本方差 S^2 的平方根 S，一是绝对平均偏差

$$d=\sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{i=1}^{n}|X_i-\overline{X}| \tag{1.40}$$

（常数 $\sqrt{\frac{\pi}{2}}$ 选择的理由是使 d 为 σ 的无偏估计，参看例 1.16 末尾处）．Fisher 赞成 S 而 Eddington 赞成 d．Fisher 在 1920 年一篇文章中谈了他的理由，其中一点是"S 包含了样本中有关 σ 的全部信息，而 d 则否"．正是在这一提法中包含了统计量的充分性这个概念。

在本节（一）中已指出：统计量的一个作用，是把样本中有用的信息集中起来．一个统计量能集中样本里的多少信息，自与统计量的具体形式有关，但也依赖于问题的统计模型（参看例 1.25 中的说明）．最好的情况是：统计量把样本中的全部信息都包含进来了．换句话说，只要算出了这个统计量的值，就是把原始样本丢掉了，也无任何损失．满足这种条件的统计量就叫做**充分统计量**．现在我们来设法给这个直观论述以严格的数学解释，从而引出充分统计量的正式定义。

记样本$(X_1, \cdots, X_n)$为X. 设有统计量$T=T(X)$. 我们可以把得出样本X的过程看成是由两步实现的：第一步观察T, 第二步在已知T的条件下去观察X. 整个样本X中所含的(有关X的样本分布的参数θ的)信息，是这两步所提供的信息的和. 第一步的信息就是统计量T所包含的信息. 因此，当且仅当第二步所提供的信息为0时，统计量才是充分的. 但第二步所提供的信息是否为0，又取决于在已知T的条件下，X的条件分布是否与参数θ无关. 因为，倘若这条件分布与θ无关，则在已知T时进一步去观察X，相当于去观察一个与θ毫无关系的量. 其中当然不包含关于θ的信息. 反之，若这条件分布与θ有关，则在已知T时观察X，还可以提供一些关于θ的信息. 这样，可以给出充分统计量的正式定义如下.

定义 1.1 设样本X的分布族为$\{F_\theta(x): \theta\in\Theta\}$, θ为分布的参数. 设$T=T(X)$为一统计量. 若在已知T的条件下，X的条件分布与θ无关，则称T是充分统计量.

在大多数情况下，样本空间$\mathscr{X}$是R^n的一部分，而统计量T可表为$T=(T_1, \cdots, T_k)$, $T_1, \cdots, T_k$都是一维变量，k一般是很小的自然数，并可以找到另一统计量$W=(W_1, \cdots, W_{n-k})$，使得变换

$$X\to(T_1(X), \cdots, T_k(X), W_1(X), \cdots, W_{n-k}(X)) \quad (1.41)$$

是一个一一对应的变换. 这时，在已知$T(X)=(t_1, \cdots, t_k)$(这等于把X限制在集合$\{X: T(X)=(t_1, \cdots, t_k)\}=B$内)的条件下，$X$与$W$有一一对应关系(确切地说，当$X$局限于$B$内时有一一对应关系)，因此，要确定在给定$T$时$X$的条件分布是否与$\theta$有关，只须确定在给定$T$时，$W$的条件分布是否与$\theta$有关. 这一事实将用于证明定理1.5.

下面考察几个例子.

例 1.18 设$X_1, \cdots, X_n\sim N(\theta, 1)$, $-\infty<\theta<\infty$, $T=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$. 作正交变换$Y_1=\sum_{j=1}^{n}\frac{1}{\sqrt{n}}X_j=\sqrt{n}\,\bar{X}=\sqrt{n}\,T$, $Y_i=$

$\sum_{j=1}^{n} a_{ij}X_j$（参看引理 1.1 的证明），此处 Y_1 实质上就是 T，而 Y_2，…，Y_n 起着 W_1，W_2，… 的作用．依引理 1.1，知 Y_1，Y_2，…，Y_n 独立，故给定 Y_1（即给定 T）时，$(Y_2, \cdots, Y_n)$ 的条件分布，就是 $(Y_2, \cdots, Y_n)$ 的无条件分布，但依定理 1.1，Y_2，…，Y_n 独立同分布，且各有分布 $N(0, 1)$．即 $(Y_2, \cdots, Y_n)$ 的分布与 θ 无关．因此证明了 $\bar{X}$ 是充分统计量．

例 1.19 回到例 1.1 和 1.2．设其中 N 已知而 M 为未知参数，考虑统计量 $T=X_1+\cdots+X_n$，即抽出的 n 个产品中的废品个数．则易见：不论是例 1.1 无放回的情况，还是例 1.2 有放回的情况，都有

$$P(X_1=x_1, \cdots, X_n=x_n | T=a) = P(X_1=x_1, \cdots, X_n=x_n, T=a)/P(T=a) = P(X_1=x_1, \cdots, X_n=x_n)/P(T=a) = 1\bigg/\binom{n}{a}.$$

当 x_1，…，x_n 都为 0 或 1，且 $x_1+\cdots+x_n=a$（否则上述条件概率为 0），在计算中用到了(1.1)、(1.2)、(1.22)和(1.23)式．由于这条件概率与参数 M 无关，证明了统计量 T 是充分的．

例 1.20 回到例 1.18．考虑统计量 $T=X_1$．这时，可以取 W_1，W_2，…为 X_2，…，X_n，在给定 T 即 X_1 的条件下，$(X_2, \cdots, X_n)$ 的条件分布即其无条件分布，它的概率密度 $\prod_{i=2}^{n}\left(\frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2}\right)$ 与 θ 有关，因此依定义 1.1，$T=X_1$ 不是充分统计量．

例 1.19 和 1.20 的结果在直观上是可以预料的．在例 1.19 中，若已知样本中的总废品数，则进一步知道“这些废品是在那几次抽得的”，对推断整批废品数 M 看来并无帮助．对例 1.20 则更明显：统计量 T 中只使用了一个观测值 X_1，把其余观测值 X_2，…，X_n 全丢了，它当然不能把样本 X_1，…，X_n 中的所有信息都集中起来．例 1.18 的结果则不能说在直观上很明显．事实上，并非在任何情况下，样本均值都能包含有关总体均值的全部信息．这

一点取决于总体分布的形式(习题 7).

直接用定义 1.1 去验证一个统计量 T 的充分性，往往要经过复杂的计算. 幸好，有下面的一般性判定法，它在应用上很方便.

定理 1.5 (**因子分解定理**) 设样本 $X=(X_1, \cdots, X_n)$的概率函数 $f_\theta(x_1, \cdots, x_n)$依赖于参数 θ, $T=T(X)$ 是一个统计量. 则 T 为充分统计量的充要条件是: $f_\theta(x_1, \cdots, x_n)$可以分解为

$$f_\theta(x_1, \cdots, x_n)=g_\theta(T(x_1, \cdots, x_n))h(x_1, \cdots, x_n) \quad (1.42)$$

的形状. 注意函数 h 不依赖于参数 θ.

这里**概率函数**的意义是: 若 X 是连续型的，则 $f_\theta(x_1, \cdots, x_n)$ 是其概率密度; 若 X 是离散型的，则 $f_\theta(x_1, \cdots, x_n)=P_\theta(X_1=x_1, \cdots, X_n=x_n)$. P_θ 表示: 有关概率是在参数值为 θ 时计算的.

这个重要的定理在二十年代就由 R. A. Fisher 提出来，它的最一般形式和严格数学证明，是 Halmos 和 Savage 在 1949 年作出的.

为确定计，考虑 X 有概率密度的情况，并补充假定 T 有$(T_1, \cdots, T_k)$的形式，且可以补充统计量 $W=(W_1, \cdots, W_{n-k})$, 适合定义 1.1 下面那一段话的条件. 以 J 记变换 (1.41) 的 Jacobi 行列式的绝对值，并将 $x_1, \cdots, x_n$ 的函数 $h(x_1, \cdots, x_n)/J$ 表为 T, W 的函数 $H(T, W)$, 则由概率密度变换公式及(1.42), 知(T, W)的密度为 $g_\theta(T)H(T, W)$. 于是得到: 在给定 T 时, W 的条件密度为

$$g_\theta(T)H(T, W)\Big/\int_{R_{n-k}} g_\theta(T)H(T, W)dW$$

$$=H(T, W)\Big/\int_{R_{n-k}} H(T, W)dW.$$

此条件密度与 θ 无关，故 T 为充分统计量. 这证明了条件(1.42)的充分性.

反之，设 T 为充分统计量，则以 $G(T, W)$记给定 T 时, W 的条件密度. 由 T 充分，知 G 与 θ 无关. 故得(T, W)的联合密度为 $g_\theta(T)G(T, W)$, 其中 $g_\theta(T)$为 T 的密度函数. 把 $G(T, W)$表

为 $x_1, \cdots, x_n$ 的函数 $h_1(x_1, \cdots, x_n)$，则仍由密度变换公式得$(X_1, \cdots, X_n)$的密度为 $g_\theta(T(x_1, \cdots, x_n))h_1(x_1, \cdots, x_n)J=g_\theta(T(x_1, \cdots, x_n))h(x_1, \cdots, x_n)$，其中 $h(x_1, \cdots, x_n)=h_1(x_1, \cdots, x_n)J$ 与 θ 无关. 这证明了条件(1.41)的必要性，定理证毕.

利用这个定理很容易判定许多重要的统计量的充分性（或不充分性），如例 1.18和 1.19.

例 1.21 设 $X_1, \cdots, X_n \sim N(a, \sigma^2)$，参数为 $\theta=(a, \sigma)$. 以 $\bar{x}$ 和 S^2 分别记样本均值和样本方差，则$(\overline{X}, S^2)$是充分统计量.

事实上，$(X_1, \cdots, X_n)$有概率密度

$$(\sqrt{2\pi}\ \sigma)^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right)$$

$$=(\sqrt{2\pi}\ \sigma)^{-n}\exp\left(-\frac{n}{2\sigma^2}(\bar{x}-a)^2\right)\cdot\exp\left(-\frac{1}{2\sigma^2}S^2\right).$$

因而有(1.42)的形式（取其中的 $h\equiv 1$）.

例 1.22 设 $X_1, \cdots, X_m \sim N(a, \sigma^2)$, $Y_1, \cdots, Y_n \sim N(b, \sigma^2)$，且两组样本独立. 记 $\overline{X}=\sum\limits_{i=1}^{m}X_i/m$, $\overline{Y}=\sum\limits_{i=1}^{n}Y_i/n$，而

$$S^2=\frac{1}{m+n-2}\left[\sum_{i=1}^{m}(X_i-\overline{X})^2+\sum_{j=1}^{n}(Y_j-\overline{Y})^2\right]. \tag{1.43}$$

则与上例类似证明：$(\overline{X}, \overline{Y}, S^2)$为充分统计量. 注意此例中两总体方差相同，而模型中的参数是 $\theta=(a, b, \sigma^2)$.

例 1.23 回到例 1.5，参数为 λ. 用条件(1.42)立即得到：$\overline{X}$ 是充分统计量.

例 1.24 设 $X_1, \cdots, X_n \sim R(0, \theta)$, $\theta>0$. 这里 $R(0, \theta)$ 记区间$(0, \theta)$上的均匀分布，记

$$T=\max(X_1, \cdots, X_n). \tag{1.44}$$

$(X_1, \cdots, X_n)$的联合密度为

$$f_\theta(x_1, \cdots, x_n)=\begin{cases}\dfrac{1}{\theta^n}, & \text{当}\ \max(x_1, \cdots, x_n)<\theta; \\ 0, & \text{其他情况}.\end{cases} \tag{1.45}$$

在本例中，样本只能取正值，故 $f_\theta(x_1, \cdots, x_n)$ 只须在集合 $\{(x_1,$

$\cdots, x_n): x_i>0,\ i=1,\ \cdots,\ n\}=B$ 上定义. (1.45)式也是在集 B 上而言. 易见, 若定义函数 $g_\theta(t)=1$ 当 $0<t<\theta$, 在其他处为 0, 而 $h\equiv 1$, 则(1.45)可改写为 $f_\theta(x_1,\ \cdots,\ x_n)=g_\theta(t)h(x_1,\ \cdots,\ x_n)$. 因此由(1.44)定义的统计量 T 是充分的.

例 1.25 设 $X_1,\ \cdots,\ X_n \sim R\left(-\frac{1}{2}+\theta,\ \frac{1}{2}+\theta\right),\ -\infty<\theta<\infty$ θ 是区间 $\left[-\frac{1}{2}+\theta,\ \frac{1}{2}+\theta\right]$的中点, 也就是总体分布的期望值. 用定理 1.5, 不难验证, 样本均值 $\bar{X}$ 并不是充分统计量(读者自证).

• 例 1.18 与例 1.25 结合, 说明了前面提到过的一件事实: 一个统计量是否为充分, 不仅取决于这统计量的形式, 也与问题的模型即样本分布族有关.

习 题

1. 一个总体有 N 个元素. 其指标值分别为 $a_1>a_2>\cdots>a_N$. 指定自然数 $M<N,\ n<N$, 并设 $m=nM/N$ 为整数. 在$(a_1,\ \cdots,\ a_M)$中不放回地随机抽出 m 个, 在$(a_{M+1},\ \cdots,\ a_N)$中不放回地随机抽出 $n-m$ 个. 写出所得样本的分布.

2. 一物体的重量 a 未知, 有两架天平可用, 其随机误差分别服从正态分布 $N(0,\ \sigma_1^2),\ N(0,\ \sigma_2^2),\ \sigma_1^2$ 和 σ_2^2 都未知. 先把该物件在第一架天平上秤两次得 $X_1,\ X_2$, 再在第二架天平上秤两次得 $X_3,\ X_4$, 然后视 $|X_1-X_2|\leqslant|X_3-X_4|$ 或否而在第一架或第二架天平上再秤 $n-4$ 次得 $X_5,\ \cdots,\ X_n$. 写出$(X_1,\ \cdots,\ X_n)$的密度.

3. 完成(1.20)式的归纳证明.

4. 利用定理 1.3 证明(1.38)式. 要把 n 取奇数和 n 取偶数的情况分开处理(提示: 先证明: 若 $X_n\leqslant Y_n\leqslant Z_n$, 而当 $n\to\infty$ 时, X_n 的分布和 Z_n 的分布收敛于同一极限, 则 Y_n 的分布也收敛于该极限).

5. 设 $T=T(X)$是充分统计量, 又 $S(X)=G(T(X))$, 而函数 $S=G(T)$ 是一对一的(即 $T_1\neq T_2\Rightarrow G(T_1)\neq G(T_2)$), 则 S 也是充分统计量.

6. 直接由充分统计量的定义出发证明: 若 $X_1,\ \cdots,\ X_n$ 是从分布族(1.8)中抽出的独立随机样本. 直接由定义出发, 证明 $\bar{X}$ 是充分统计量(提示: 用例 1.18 的方法, 作线性变换 $Y_1=\sum_1^m X_i,\ Y_i=X_i,\ i=2,\ \cdots,\ n$).

7．设 $X_1, \cdots, X_n \sim N(\theta, \theta^2)$, $\theta>0$. 问 $\overline{X}$ 是否仍为充分统计量?（提示:用因子分解定理）

8．证明 $P(X_{(m)}<x)$ 的积分表达式(1.24)(提示: 为证 $\sum_{i=m}^{n}\binom{n}{i}p^i(1-p)^{n-i}=m\binom{n}{m}\int_0^p t^{m-1}(1-t)^{n-m}dt$, 注意 $p=0$ 时两边相等，两边对 p 的导数也一样).

9．仔细写出推导(1.26)的过程(提示:先证明

$$\int\cdots\int_{a<x_1<x_2<\cdots<x_n<b} f(x_1)\cdots f(x_n)dx_1\cdots dx_n=\frac{1}{n!}[F(b)-F(a)]^n).$$

10．设 $\xi\sim\chi_n^2$. a 为常数，计算 $E\xi^a$ 和 $\mathrm{Var}(\xi^a)$. 注意 a 取那些值时这些量才存在．又问: ξ 的密度的最大值在那一点达到?

11．设 $\xi\sim\chi_n^2$. 证明当 $n\to\infty$ 时，$(\xi-n)/\sqrt{2n}\xrightarrow{\mathscr{L}}N(0,1)$. 利用这个事实，给出 χ_n^2 的 p 分位点与 $N(0, 1)$ 的 p 分位点之间的一个近似关系.

12．设 $\xi\sim t_n$. 计算 $\mathrm{Var}(\xi_n)$. 注意 n 取何值时，$\mathrm{Var}(\xi_n)$才存在.

13．设 $X_1, \cdots, X_n\sim N(a, \sigma^2)$. $\overline{X}=\sum_1^n X_i/n$, $S^2=\sum_1^n(X_i-\overline{X})^2/(n-1)$. 记 $\xi=(X_1-\overline{X})/S$. 找出 ξ 与 t 分布的联系，因而定出 ξ 的密度(提示: 作正交变换 $Y_1=\sqrt{n}\ \overline{X}$, $Y_2=\sqrt{\frac{n}{n-1}}(X_1-\overline{X})$, $Y_i=\sum_{j=1}^n c_{ij}X_j$, $i=3, \cdots, n$, 利用定理1.1).

14．非中心 χ^2 变量 $\xi=\sum_1^n(X_i+a_i)^2$, $X_1, \cdots, X_n\sim N(0, 1)$, $a_1, \cdots, a_n$ 为常数．证明: ξ 的分布只依赖于 n 和 $\delta=(\sum_1^n a_i^2)^{1/2}$ (提示: 作正交变换，使 $Y_1=\sum_{j=1}^n a_jX_j/\delta$).

15．设 $X_1, \cdots, X_n$ 独立，$X_i\sim N(0, \sigma_i^2)$, $i=1, \cdots, n$. 定义 $\xi=\sum_1^n(X_i-Z)^2/\sigma_i^2$, 其中 $Z=\sum_1^n\frac{X_i}{\sigma_i^2}\Big/\sum_1^n\frac{1}{\sigma_i^2}$, 求 ξ 的分布(提示: 作适当的正交变换).

16．设 $X_1, \cdots, X_n\sim N(a, \sigma^2)$, $\overline{X}$ 为样本均值, $\xi=f(X_1, \cdots, X_n)$ 满足条件 $f(X_1+c, \cdots, X_n+c)=f(X_1, \cdots, X_n)$, 对任何常数 c. 证明: ξ 与 $\overline{X}$ 独立(提示: 作定理1.1中的变换，把 ξ 表为 $Y_1, \cdots, Y_n$ 的函数，证明此函数只依赖于 $Y_2, \cdots, Y_n$). 举出此结果的一些应用(样本方差，极差).

第二章　点　估　计

这个题目在上一章中已经多次提到过了．设样本 X 的分布依赖于参数 θ, θ 取值于 Θ 内，Θ 称为参数空间．设 $g(\theta)$是一个定义在 Θ 上的已知函数．要通过样本 X 去估计 $g(\theta)$的值．这里 θ 和 $g(\theta)$都可以是一维或者多维，一个重要的特例是 $g(\theta)=\theta$，或 $g(\theta)$等于 θ 的某个分量．例如，正态分布 $N(a, \sigma^2)$有参数 $\theta=(a, \sigma^2)$，我们只估计 $g(\theta)=a$．在本书所讨论的多数情况下，样本 X 有$(X_1, \cdots, X_n)$的形式，其中 $X_1, \cdots, X_n$ 相互独立相同分布，这时 θ 是总体分布的参数，而我们可用估计总体分布的参数这个提法．一般统计著作中多用这个提法．然而应当注意：如在§1.2(四)中所指出的，只有在简单随机样本的情况下，"总体分布"一词才有明确的意义．或更确切地说，才能由总体分布决定样本分布．直接从样本分布出发就没有这种问题．

$g(\theta)$常称为"待估函数"或"被估计量"．为估计它，就要决定一个适当的函数 $\hat{g}$，定义于样本空间 $\mathscr{X}$，使得每当有了样本 X，就以 $\hat{g}(X)$作为 $g(\theta)$的估计值．$\hat{g}(X)$是一个统计量，因其用于作点估计，常称之为**估计量**．

点估计是数理统计学中内容很丰富的一个分支．它主要包括制定估计量的一般方法，制定有关估计量优良性的种种合理准则，寻求某种特定准则下的最优估计量，以及证明某一特定的估计量(用直观方法或某种一般性方法得到的)在某种准则之下有最优性等等．本书有关点估计的内容，除一部分(基础的和一般性的)内容集中在本章讨论外，还在第五～七章中占一定的地位．

§2.1　矩估计与极大似然估计

(一) 矩估计方法

定义 2.1　设有样本 $X_1, \cdots, X_n$，而 k 为自然数，则

$$a_{nk}=\frac{1}{n}\sum_{i=1}^{n}X_i^k \quad 及 \quad m_{nk}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^k \tag{2.1}$$

分别称为 **k 阶样本原点矩**和 **k 阶样本中心矩**, 这里 $\overline{X}_n=a_{n1}=\sum_{i=1}^{n}X_i/n$ 是样本均值.

当 $k=1$ 时 $m_{nk}=0$, 又 m_{n2} 与样本方差 S^2 只相差一常数因子(见(1.18)式). 一般,样本中心矩可通过样本原点矩表出:

$$m_{nk}=\sum_{r=0}^{k}(-1)^{k-r}a_{nr}a_{n1}^{k-r} \quad (a_{n0}=1). \tag{2.2}$$

设 $X_1, \cdots, X_n$ 是简单随机样本(即 $X_1, \cdots, X_n$ 独立同分布,见(1.10)式前面的说明), 则有总体分布 F. 这时, 样本矩可用于估计分布 F 的相应阶的矩(称为**总体矩**). 即若以 α_k 和 μ_k 分别记 F 的 k 阶原点矩及 k 阶中心矩, 则可用 a_{nk} 估计 α_k, m_{nk} 估计 μ_k. 这是一种基于直观的方法. 它的一个依据是: a_{nk} 是 α_k 的无偏估计(见(1.35)式的说明)

$$Ea_{nk}=\frac{1}{n}\sum_{i=1}^{n}EX_i^k=\frac{1}{n}\sum_{i=1}^{n}\alpha_k=\alpha_k. \tag{2.3}$$

但如用 m_{nk} 估计 μ_k, 则一般不是无偏的, 不过当样本大小 n 较大时,偏差不显著,且必要时可作一些修正,使之成为无偏的.

例 2.1 μ_2 是总体分布的方差. 若用 m_{n2} 估计 μ_2,则

$$\begin{aligned}Em_{n2}&=\frac{1}{n}\sum_{i=1}^{n}EX_i^2-E(\overline{X}_n^2)\\&=\mu_2+\alpha_1^2-(\mathrm{Var}(\overline{X}_n)+(E\overline{X}_n)^2)\\&=\mu_2+\alpha_1^2-\left(\frac{1}{n}\mu_2+\alpha_1^2\right)=\frac{n-1}{n}\mu_2.\end{aligned} \tag{2.4}$$

因而 m_{n2} 不是 μ_2 的无偏估计, 且是系统地偏低. 在本例中修正不难,只须用

$$\frac{n}{n-1}m_{n2}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2=S^2$$

代替 m_{n2},就得到 $ES^2=\mu_2$, 即 S^2 是总体方差 μ_2 的无偏估计. 这正是(1.18)式定义样本方差时, 采用因子 $\frac{1}{n-1}$ 的原因. 当 n 较

大时，二者差别不大.

当 $k\geqslant4$ 时，就不能通过这样简单的修正，由 m_{nk} 得出 μ_k 的无偏估计，而需要使用几个中心矩. 一般在应用上则不去作任何修正，而容忍一些偏差存在. 如上所述，对较大的 n 这与应用无损.

现设总体分布依赖于参数 θ, 而待估函数 $g(\theta)$(假定是一维的)可表为总体分布的若干个矩的函数：

$$g(\theta)=G(\alpha_1, \cdots, \alpha_k, \mu_2, \cdots, \mu_l). \tag{2.5}$$

且有了简单随机样本 $X_1, \cdots, X_n$. 则可以用下面的方法构造出 $g(\theta)$的一个估计量：用 a_{ni} 估计 α_i, m_{ni} 估计 μ_i, 然后令($X=(X_1, \cdots, X_n)$)

$$\hat{g}(X)=G(a_{n1}, \cdots, a_{nk}, m_{n2}, \cdots, m_{nl}), \tag{2.6}$$

即以之作为 $g(\theta)$的估计，(2.6)称为$g(\theta)$的一个**矩估计**. 需要注意的是：同一个$g(\theta)$往往可以有几种不同的方法表为(2.5)的形式. 因此，矩估计也可以有许多. 这样，我们最好把矩估计看成构造估计量的一种一般性方法，而不看成是一种有固定程序的算法.

例 2.2 设总体分布为 Poisson 分布

$$P_\theta(X=x)=\frac{1}{x!}e^{-\theta}\theta^x, \ x=0, 1, 2, \cdots, 0<\theta<\infty, \tag{2.7}$$

θ 为未知参数(注意，(2.7)式中的 X 并非样本，而是总体变量，见(1.11)式下面的说明). 设 $X_1, \cdots, X_n$ 为抽自此总体的简单随机样本，要估计 $g(\theta)=\theta$.

如所周知，对 Poisson 分布(2.7)，参数 θ 既是其均值 α_1, 又是其方差 μ_2. 按二者可分别构造其矩估计 a_{n1} 和 μ_{n2}, 它们自然是不同的.

在(2.7)式中，表示概率的记号 P 下面加了一个足标 θ, 它表示在计算概率时，是从分布中的参数值为 θ 出发的. 以后将要出现的记号 E_θ, Var_θ 等都是这个意义. 这一点在统计中很重要. 因为在统计中，样本或总体的可能分布不是一个，而是一个族. 在计算与这种分布有关的量时，明确指出其参数值常是很重要的，不然

就可能引起混淆.

下面再举几个例子说明求矩估计的方法.

例 2.3 设 $X_1, \cdots, X_n \sim R(\theta_1, \theta_2)$, 参数 $\theta=(\theta_1, \theta_2)$, 在 $-\infty<\theta_1<\theta_2<\infty$ 内变化, 求 θ_1 和 θ_2 的矩估计.

易算出总体分布的均值为 $\alpha_1=\frac{1}{2}(\theta_1+\theta_2)$, 方差为 $\frac{1}{12}(\theta_2-\theta_1)^2$. 分别以 $\bar{X}$ 和 S^2 记样本均值和样本方差, $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 记 θ_1 和 θ_2 的矩估计, 按矩估计法, $(\hat{\theta}_1, \hat{\theta}_2)$ 应是方程组 $\{\bar{X}=(\hat{\theta}_1+\hat{\theta}_2)/2, S^2=(\hat{\theta}_2-\hat{\theta}_1)^2/12\}$ 的解, 故得

$$\hat{\theta}_1=\bar{X}-\sqrt{3}S, \quad \hat{\theta}_2=\bar{X}+\sqrt{3}S. \tag{2.8}$$

若要估计 $\theta_2-\theta_1$, 即分布区间之长, 可用 $\hat{\theta}_2-\hat{\theta}_1=2\sqrt{3}S$.

例 2.4 设总体分布有概率密度

$$f(x, \theta)=f(x, \theta_1, \theta_2)$$
$$=\begin{cases} \dfrac{\theta_2}{\Gamma\left(\dfrac{1+\theta_1}{\theta_2}\right)} x^{\theta_1}\exp(-x^{\theta_2}), & x>0; \\ 0, & x\leqslant 0. \end{cases} \tag{2.9}$$

参数 $\theta=(\theta_1, \theta_2)$ 的变化范围为 $-1<\theta_1<\infty$, $\theta_2>0$. 设 $X_1, \cdots, X_n$ 为抽自此总体的简单随机样本, 求 θ_1 和 θ_2 的矩估计.

简单计算得到总体分布的前两阶矩为:

$$\alpha_1=\Gamma\left(\frac{2+\theta_1}{\theta_2}\right)\Big/\Gamma\left(\frac{1+\theta_1}{\theta_2}\right),$$
$$\alpha_2=\Gamma\left(\frac{3+\theta_1}{\theta_2}\right)\Big/\Gamma\left(\frac{1+\theta_1}{\theta_2}\right). \tag{2.10}$$

按矩估计作法, 用 a_{n1} 和 a_{n2} 分别代替 (2.10) 中的 α_1 和 α_2, $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 分别代替 θ_1 和 θ_2, 即得出一个方程组, 其解就是 θ_1, θ_2 的一个矩估计. 但此处得不出简单的解析表达式, 而只能用数值方法.

例 2.5 以 μ_k 记总体分布的 k 阶中心矩, α_1 记其均值, 则

$$\beta_1=\mu_3/\mu_2^{3/2}, \quad \beta_2=\mu_4/\mu_2^2, \quad v=\sqrt{\mu_2}/\alpha_1$$

分别称为总体分布的偏度、峰度和变异系数. 当总体分布关于某点对称时, $\beta_1=0$. 故 β_1 是作为总体分布的"非对称性"或"偏倚

性"的一种度量．特别，当总体分布为正态时，$\beta_1=0$ 而 $\beta_2=3$．因此，β_1 与 0 的差异及 β_2 与 3 的差异，可作为总体分布与正态的偏离的一种度量．变异系数 v 是衡量总体分布散布程度的一项指标，但是这散布程度是以总体均值为单位来度量．用矩估计法，直接作出这些量的估计．

另有一类直接由矩定义的，在较早期统计文献中常提到的量，叫**半不变量**．其定义如下：设总体分布有直到 k 阶为止的原点矩 $\alpha_1,\ \cdots,\ \alpha_k$，将其特征函数 $\varphi(t)$ 在 $t=0$ 的附近表为

$$\varphi(t)=1+\sum_{j=0}^{k}\frac{1}{j!}\alpha_j(it)^j+o(t^k).$$

两边取对数，再作一些简单的代数整理，易得

$$\log\varphi(t)=1+\sum_{j=1}^{k}\frac{1}{j!}x_j(it)^j+o(t^k). \tag{2.11}$$

系数 x_j 就称为总体分布的 j 阶半不变量，x_j 可通过总体的前 j 阶原点或中心矩表出，例如

$$x_1=\alpha_1,\ x_2=\mu_2,\ x_3=\mu_3,\ x_4=\mu_4-3\mu_2^2,\ x_5=\mu_5-10\mu_2\mu_3 \tag{2.12}$$

等等．这些量的矩估计都可以直接写出．

例 2.6 矩估计法也可用于多维样本．例如，设 $(X_i,\ Y_i)$，$i=1,\ \cdots,\ n$ 为从一个二维总体中抽出的简单随机样本．总体分布的协方差和相关系数，分别用 μ_{11} 和 P 记，则 μ_{11} 和 ρ 可分别用

$$m_{11}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\left(\overline{X}=\sum_{i=1}^{n}X_i/n,\ \overline{Y}=\sum_{i=1}^{n}Y_i/n\right) \tag{2.13}$$

和

$$r=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\Big/\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)^{1/2} \tag{2.14}$$

去估计，m_{11} 和 r 分别称为**样本协方差**和**样本相关系数**，详见第七章．

(二) 极大似然估计

定义 2.2 设样本 X(不一定是简单随机样本)有概率函数 $f(x, \theta)$(概率函数的意义, 见(1.42)式下面的说明). 这里 θ 为参数,在参数空间 Θ 内取值. 当固定 x 而把 $f(x, \theta)$看成是 θ 的定义在 Θ 上的函数时,它称为**似然函数**.

所以, 概率函数与似然函数可以说是一回事, 只是看法不同: 前者是固定 θ 而看成是 x 在样本空间$\mathscr{X}$上的函数,后者则固定 x 而看成是 θ 在 Θ 的函数. 这个差别在统计上的意义如下: 不妨把参数 θ 和样本 x 分别看成是"原因"和"结果": 定了 θ 的值,就完全确定了样本分布,也就定下了得到种种结果(x)的机会大小. 这是从正面看. 从反面看, 当有了结果(样本)x 时, 我们问: 当参数 θ 取各种不同的值(原因时,导出这个结果(x)的可能性有多大?这个问题的回答引出似然函数的概念. "似然"的字面意义就是"看起来象". 说仔细一些,就是: 当我们有了结果 x 时,这结果看来是由原因 θ 而产生的可能性,与似然函数值 $f(x, \theta)$成比例. 由于统计推断是由样本推断参数, 这个看法就可以作为一种统计推断方法的哲理基础. 事实上, 确有一些统计学家作这样的主张. 它们把基于每个参数值的"似然性"去进行统计推断这个原则, 叫做**似然原则**. 应当注意的是: 反映 θ 的似然性的 $f(x, \theta)$, 虽然是源出于一个概率论概念——概率函数,本身并不是通常意义下的概率,它自然也没有频率解释.

似然原则的一项重要应用, 就是由下述定义所确定的极大似然估计.

定义 2.3 在定义 2.2 的记号下, 若 $\hat{\theta}(X)$是一个统计量, 满足条件

$$f(x, \hat{\theta}(x))=\sup_{\theta\in\Theta} f(x, \theta), \ x\in\mathscr{X}. \tag{2.15}$$

$\mathscr{X}$ 是样本空间,则称 $\hat{\theta}(X)$是 θ 的**极大似然估计**. 若待估函数是 $g(\theta)$,则称 $g(\hat{\theta}(X))$为极大似然估计.

按我们上面对"似然性"一词所作的解释，θ 的极大似然估计 $\hat{\theta}(x)$，就是在已得样本 x 的情况下，似然性最大的那个 θ 值．知道 x 自然不能唯一决定 θ，取那个值去估计 θ 呢？要取那个"看起来最象"的值．如果我们同意上述关于似然性的直观解释，这个选择不仅合乎逻辑，甚至可以说是唯一合理的．

$\hat{\theta}$ 的确定要解一个极值问题．有时这是很困难的，而不能不采用数值方法．在 $X=(X_1, \cdots, X_n)$ 为简单随机样本而总体分布有概率函数 $f_\theta(x)$ 时，似然函数为 $f(x, \theta)=\prod_{i=1}^{n} f_\theta(x_i)$．这时，$f(x, \theta)$ 的对数

$$\log f(x, \theta)=\sum_{i=1}^{n} \log f_\theta(x_i) \tag{2.16}$$

在使用上较方便．$\log f(x, \theta)$ 称为**对数似然函数**．自然，极大似然估计 $\hat{\theta}(x)$ 也可以等价地定义为

$$\log f(x, \hat{\theta}(x))=\sup_{\theta\in\Theta} \log f(x, \theta). \tag{2.17}$$

例 2.7 前面讨论过的一些例子中，有些其极大似然估计很易求得．例如，据(1.1)或(1.2)式易知，在例 1.1 和例 1.2 之下，批中废品数 M 的极大似然估计都是 $\frac{N}{n}(X_1+\cdots+X_n)=N\bar{X}$，或者说，批废品率 $\frac{M}{N}$ 的极大似然估计，就是所抽样本中的废品率 $\bar{X}$．在例 1.4 中，若设 σ 已知，则对数似然函数为 $n\log\frac{1}{\sqrt{2\pi}}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2$．由此导出 a 的极大似然估计为 $\bar{X}$．例 1.5 的对数似然函数为 $n\log\lambda-\lambda\sum_{i=1}^{n}x_i$，导出 λ 的极大似然估计为 $1/\bar{X}$．在这几个例子中，极大似然估计同时也是矩估计．

例 2.8 设 $X_1, \cdots, X_n \sim N(a, \sigma^2)$，参数 $\theta=(a, \sigma)$．求 θ 的极大似然估计．在本例中，对数似然函数为 $n\log\frac{1}{\sqrt{2\pi}}-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2$．通过求偏导数并命之为 0，得出方程

$$-n/\sigma+\sum_{i=1}^{n}(x_i-a)^2/\sigma^3=0,\quad \sum_{i=1}^{n}(x_i-a)/\sigma^2=0. \tag{2.18}$$

象(2.18)这样，通过对似然函数或对数似然函数求导而得的方程，称为**似然方程**. 在f可导而$f(x, \theta)$在Θ上的最大值点在Θ的内部时，它必然是似然方程的根. 但由于我们不一定在每一具体场合都明确这些条件是否成立，且似然方程也可以不止一个根，因此往往还需要经过验证，才能肯定似然方程的某一解是(或不是)极大似然估计. 对本例而言，易算出(2.18)有唯一解

$$\hat{a}=\overline{X},\ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=m_{n2}, \tag{2.19}$$

且不难验证它确是极大似然估计(习题1).

例2.9 具有概率密度函数

$$f_\theta(x)=\frac{1}{\pi[1+(x-\theta)^2]},\ -\infty<x<\infty \tag{2.20}$$

的分布称为**Cauchy分布**. 参数θ可以取任意实数值.

现设$X_1, \cdots, X_n$为抽自具Cauchy分布的总体的简单随机样本，要求θ的极大似然估计. 本例似然方程为

$$\sum_{i=1}^{n}\frac{x_i-\theta}{1+(x_i-\theta)^2}=0. \tag{2.21}$$

这方程只能用数值解，比方说，用牛顿法，而以样本中位数作为初始值. 注意，分布(2.20)关于θ对称，故θ就是总体分布中位数. 这个分布不存在期望值，因此不能用矩法估计θ.

例2.10 设简单随机样本$1°\sim R(0, \theta)$, $\theta>0$; 或$2°\sim R(\theta_1, \theta_2)$, $-\infty<\theta_1<\theta_2<\infty$. 写出这两个情况下的似然函数，直接看出极大似然估计为: $1°\ \hat{\theta}=\max(X_1, \cdots, X_n)$, $2°\ \hat{\theta}_1=\min(X_1, \cdots, X_n)$, $\hat{\theta}_2=\max(X_1, \cdots, X_n)$. 本例似然函数不连续，不能用似然方程求解.

例2.11 再考虑例1.5. 仍设元件寿命的分布为(1.8)，但把试验方式作如下的修改：指定一个时刻$T>0$. 试验进行到全部抽出的n个元件都失效，或到时刻T为止. 记X_i=元件i的寿命或T，视元件在时刻T时已失效或否而定. 要由$X_1, \cdots, X_n$作λ的极大似然估计.

若$x_i<T$，则总体分布在x_i点的密度为$\lambda e^{-\lambda x_i}$. 若$x_i=T$，则

表示"元件 i 的寿命 $\geqslant T$，其概率为 $\int_T^{\infty}\lambda e^{-\lambda x}dx=e^{-\lambda T}$. 故单个样本 x_i 的概率函数，应为 $\lambda e^{-\lambda x_i}$ 或 $e^{-\lambda T}$，视 $x_i<T$ 或否而定. 现为方便计，设前 r 个元件在时刻 T 前已失效，而后 $n-r$ 个则否(必要时这可通过调整样本的编号达到). 则似然函数为

$$f(x, \lambda)=\lambda^r e^{-\lambda(x_1+\cdots+x_r)}e^{-\lambda(n-r)T},$$

似然方程的根为

$$\hat{\lambda}=\frac{1}{r}[X_1+\cdots+X_r+(n-r)T].$$

易验证这确是 λ 的极大似然估计.

本例中的试验叫**定时截尾试验**. 另一种方式是先定下一个自然数 r，待有 r 个元件用坏时试验即停止，这叫做**定数截尾试验**. 也可以两种截尾都有. 本例一个有趣之处在于：它的概率函数是"混合式"的：既不全然是概率密度，也不全然是离散概率.

例 2.12 设有 k 个事件 $A_1, \cdots, A_k$ 两两互斥，其概率 $p_1, \cdots, p_k$ 之和为 1. 将试验独立地重复 n 次，以 X_i 记 A_i 发生的次数，$i=1, \cdots, n$. 求 $p_1, \cdots, p_k$ 的极大似然估计.

此处 $(X_1, \cdots, X_k)$ 服从多项分布 $M(n; p_1, \cdots, p_k)$，其概率函数为 $\frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}$. 取对数并在 $p_1+\cdots+p_k=1$ 的条件下求极值，易算出极值点为

$$\hat{p}_i=X_i/n, \ i=1, \cdots, k.$$

且不难验证这确是极大似然估计，当 $k=2$ 时，得到二项分布的情形.

在有些问题中，$p_1, \cdots, p_k$ 都是另一些参数 $\theta_1, \cdots, \theta_r$ 的函数，$r\leqslant k$. 这时，去掉与 $\theta_1, \cdots, \theta_r$ 无关的部分后，得对数似然函数为 $\sum_{i=1}^{k}x_i\log p_i(\theta_1, \cdots, \theta_r)$. 通过建立似然方程找出 $\theta_1, \cdots, \theta_r$ 的极值点 $\hat{\theta}_1, \cdots, \hat{\theta}_r$，若能验明这确是最大值点，则得到 p_i 的极大似然估计为 $p_i(\hat{\theta}_1, \cdots, \hat{\theta}_r)$，这一般与(2.23)不一样. 一个实际例子如下：人的血型有 O, A, B, AB 四种，决定血型的基因有 A、B、O 三种.

子代的血型由父母各出一基因构成：OO 为 O 型，AA 和 AO 为 A 型，BB 和 BO 为 B 型，AB 为 AB 型（AO 表示：父母之一提供基因 A，另一提供 O．余类推）．在特定的一群人（如某一地区某一人种）中，基因 A、B、O 的频率 θ_1, θ_2, θ_3 无法直接观察，而一个人的血型则可测定．以 p_1, p_2, p_3, p_4 分别记 O、A、B、AB 这四种血型的频率，则应有（注意 $\theta_1+\theta_2+\theta_3=1$）

$$p_1=\theta_3^2,\ p_2=2\theta_1\theta_3+\theta_1^2,\ p_3=2\theta_2\theta_3+\theta_2^2,\ p_4=2\theta_1\theta_2. \tag{2.22}$$

从这一群人中随机地抽出 n 个（假定这一群人数相对于 n 很大，以至抽样可近似认为是有放回的），分别以 n_1, n_2, n_3 和 n_4 记其中具血型 O、A、B、AB 的人数，则为得 θ_1, θ_2, θ_3 的极大似然估计，要找

$$\theta_3^{2n_1}(2\theta_1\theta_3+\theta_1^2)^{n_2}(2\theta_2\theta_3+\theta_2^2)^{n_3}(2\theta_1\theta_2)^{n_4} \tag{2.23}$$

在 $\theta_1+\theta_2+\theta_3=1$ 的约束下的极值点 $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$．这只能用数值方法求解．

矩估计和极大似然估计是两种基本的点估计方法．在 §2.3 中将看到，从大样本的观点看，极大似然估计一般优于点估计．以此之故，它受到更大的重视．但矩估计也有其优点．一则它对样本分布形式要求少：在极大似然估计的情况，概率函数须有简单的解析表达式，参数 θ 也需取值于欧氏空间．对矩估计则无这类限制，如例 2.5．另外，在矩估计法能用的场合，其计算一般比极大似然估计简单些．

（三）若干历史情况

矩估计法最初是由 K. Pearson 在 1894 年一项工作中提出来的．在 1894～1902 年期间他发表了一系列工作，涉及这个方法．其中最重要的是 1902 年发表在 Biometrika 上的文章．Pearson 关于矩估计的工作是与他所提出的一个著名的分布族-Pearson 分布族密切相关．

在上世纪很长一段时期，有一些学者，包括生物学家兼统计学家 F. Galton，认为在实际问题中出现的数据，基本上都可以用正

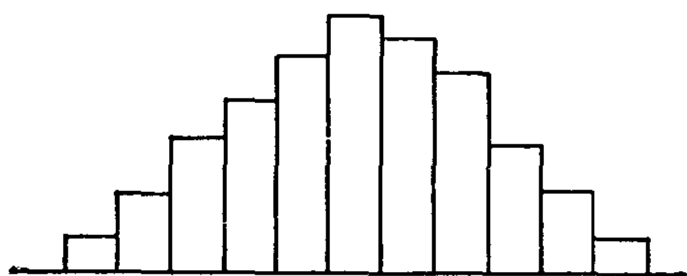
态分布来描述. Pearson 考察了一些生物学方面的数据，发现不少有显著的偏倚，而不适合用正态分布(它是对称的)去描述. 他从某些一般性考虑出发，提出了一个微分方程

$$\frac{dy}{dx}=\frac{y(x-a)}{bx^2+cx+d}. \tag{2 24}$$

而把适合这方程的所有概率密度函数 $y(x)$ 挑选出来，构成一个分布族. 后人将之称为 **Pearson 分布族**，它含有四个参数 a、b、c、d. 这是一个颇大的分布族：正态分布，指数分布以及 x^2、t、F 等重要分布都属于这个族，整个族又分为若干个子族-类型. 例如，在极值统计中重要的 **Gamma 分布**，就属于 Pearson III 型.

Pearson 认为，这个分布族足够用以描述生物学上的数据. 问题在于对一组特定的数据，要适当选择参数 a、b、c、d. Pearson 注意到：这些参数可表为密度 y 的前四阶矩 $\alpha_i=\int_{-\infty}^{\infty}y(x)x^i dx$，$i=1, \cdots, 4$ 的函数. 因此他提出用样本矩 a_{ni} 估计 α_i 并代入这些函数，就可得到 a, b, c, d 的估计. 不过，Pearson 实际上不是把这当成一个方法直接提出来，他是从“曲线拟合”的观点来处理问题，矩估计方法是他由此所作出的结论.

要说清楚 Pearson 的想法，须得介绍一下一个常见的统计概念——**直方图**. 设有了观察数据即样本 $X_1, \cdots, X_n$. 选择两个适当的常数 g、h, $h>0$. 把$(-\infty, \infty)$分为一些小区间 $\Delta_i=[g+(i-1)h, g+ih)$, $i=0, \pm1, \pm2, \cdots$, 以 n_i 记 $X_1, \cdots, X_n$ 中落在 Δ_i 内的个数. 以 Δ_i 为底，$n_i/(nh)$为高作一矩形，$i=0, \pm1, \pm2, \cdots$，就得到一个如右边那种形状的图形，这就是直方图. 它定义的函数 $\hat{y}(x)=n_i/(nh)$，当 $x\in\Delta_i$ 可以作为总体分布密度函数的一个估计：因为总共作了 n 次观察，落在 Δ_i 内的有 n_i 次，故总体分布在



Δ_i 内的概率，可以用 n_i/n 去估计．而 $n_i/(nh)$ 就是 Δ_i 内单位长的概率(即密度)的估计．Pearson 用最小二乘准则：确定参数 a、b、c、d，使由(2.24)确定的密度 $y(x)$，使 $\int_{-\infty}^{\infty}(y(x)-\hat{y}(x))^2dx$ 达到最小．Pearson 在其 1902 年文章中证明：若略去某些高次项不计，则这个最小二乘问题的解为：取 a、b、c、d，使 $\int_{-\infty}^{\infty}y(x)x^i dx(=\alpha_i)=i$ 阶样本矩 a_{ni}，$i=1, \cdots, 4$．

在现代数理统计学中，对Pearson 分布族的重要性的估计，不象当初 Pearson 设想的那么高．但是他在这些工作中所发展的矩估计法，则一直被当作一个普遍有用的点估计方法而受到重视．

极大似然估计是 R. A. Fisher 在 1912 年的一项工作中提出来的．在正态分布这个特殊情况下，这方法可追溯到 Gauss 在上世纪初关于最小二乘法的工作．Fisher 在上述 1912 年工作中，批评了矩法和最小二乘法，提出了极大似然估计法．

Fisher 走的是与 Pearson 不同的路线．他认为，统计问题模型的选定不是一个数学问题，而要根据实际．他是在假定已选好模型 $f(x, \theta)$ 的条件下，去考虑参数 θ 的估计问题．这个提法与现代数理统计中点估计问题的提法一致．这个思想他在以后一些工作，特别是前面多次提到的他的 1922 年工作中，作了发挥．他提出极大似然估计法所依据的想法，基本上就是我们在定义似然函数时所解释的．

Fisher 很重视极大似然估计．在上述 1922 年工作中，尤其是在 1925 年发表的«Theory of statistical estimation»一文中，对这一估计作了许多研究．顺便提一句：上述 1925 年工作，一般认为是近代点估计理论的奠基性工作．

另一个要注意之点是，Fisher 更多地是从"极大似然估计能集中样本里的多少信息"这个角度，去研究这个估计的．那一个时期，他已提出了充分统计量的概念．有一段他相信，极大似然估计量必是充分统计量，因而在"集中信息"这个角度看是优越的．后来他发现了这一点不对．现在我们很容易举出反例(习题 5)．不

过，在这些研究中所发展的一些概念，如 Fisher 信息量等，在以后的工作中有很大的影响。且不论极大似然估计在实用上的极端重要性，就以它在理论上的促进作用来看，也是极显著的。可以说，由 Fisher 开始的工作，直到现在仍是热门的研究课题。

在一般统计教本和著作中，大抵都有这样的看法：作为一种估计方法，极大似然估计优于矩估计。但有的学者也指出：在一般地肯定这一点的同时，也应注意到：二者的出发点，即所要解决的问题，多少有些不同：Pearson 要解决的，是用最小二乘法作曲线拟合的问题。作为这一问题的解(在 Pearson 曲线族内)，矩法优于极大似然法。

§2.2 无偏估计

(一) 无偏估计和一致最小方差无偏估计

无偏估计的概念，前在例 1.16 (结尾处)、例 2.1 及其他几处地方都已提到过了。现在再将其正式定义表述如下：

定义 2.4 设样本 X 的分布依赖参数 θ, θ 在参数空间 Θ 内取值，$g(\theta)$是定义于 Θ 上的已知函数(取实数或实向量为值)，$\hat{g}(X)$是 $g(\theta)$的一个估计量。如果

$$E_\theta\hat{g}(X)=g(\theta), \text{对任何 } \theta\in\Theta, \tag{2.25}$$

则称 $\hat{g}(X)$为 $g(\theta)$的一个无偏估计(符号 E_θ 的解释参看例 2.2)。

估计的无偏性这个概念，体现了我们在§1.3 (二)中提到的"频率学派"或"古典学派"的思想。即这个概念只有在大量重复下才有意义。设想这样一个情况：每天作抽样以对 $g(\theta)$进行估计。第 i 天的样本为 $X^{(i)}$，估计值为 $\hat{g}(X^{(i)})$，一共作了 n 天。设 $X^{(1)}$，…，$X^{(n)}$是独立同分布的，则如 $\hat{g}$ 有无偏性，按大数定律，以概率 1 当 $n\to\infty$ 时有

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}(X^{(i)})\to E_\theta\hat{g}(X)=g(\theta).$$

就是说，尽管一次估计的结果 $\hat{g}(X^{(i)})$不一定恰好等于 $g(\theta)$，但在

大量重复使用时，多次估计的算术平均值，可以任意接近被估计值 $g(\theta)$. 如果这估计量 $\hat{g}$ 只用一次，则无偏性这个概念实际上不说明什么问题. 因为 $\hat{g}$ 为无偏并不保证在任何情况下(即对任何样本)，估计值 $\hat{g}(X)$ 必重合于 $g(\theta)$. 无偏性只是保证 $\hat{g}$ 没有系统误差，即用 $\hat{g}$ 估计 $g(\theta)$ 时，不是系统地偏高或偏低：它可以有偏差，但偏差是随机性的，有时大于 0，有时小于 0，而平均为 0. 问题在于这"平均为 0"一点，只有在大量重复时才能体现出来.

因此，估计的无偏性对现实问题的意义如何，必须根据这问题的具体情况去考察. 举两个例子. 设一商店每日从一工厂中取货一批. 每次取货时，从批中随机抽出一些产品以对其废品率 p 作估计，设每批有 N 件，每件单价为 a 元. 双方议定：若某日 p 的估计值为 $\hat{p}$，则商店付给工厂 $N(1-\hat{p})a$ 元. 这时，对一日的情况而言，$\hat{p}$ 可能偏高或偏低，因而有一方要吃一点亏. 但从长远看，若 $\hat{p}$ 是 p 的无偏估计，则平均说来那一方都不吃亏. 在这个例子中，无偏性无疑是一个很有用、很合理的准则.

现在设想这样一个情况：某工厂每周进原料一批. 在投入使用前，由工厂实验室对原料中某种成分含量的百分率 p 作一估计. 根据估计值 $\hat{p}$ 采取相应的工艺调整措施. 无论 $\hat{p}$ 在那一次比真正的 p 偏高或偏低，都会使所采取的措施不理想而有损于产品质量. 在此，纵使 $\hat{p}$ 是 p 的无偏估计，在长期使用中，各周的估计偏差也不能正负抵消. 因此在本例中，估计 $\hat{p}$ 的无偏性就没有多大实际意义.

不过，在目前点估计的理论和实用中，无偏性仍占据很重要的地位. 除了历史的因素外，还有两个原因. 一是无偏性的要求只涉及一阶矩(均值)，在数学上处理较方便. 另外，在没有其他合理准则可循时，人们心理上觉得：一个有无偏性的估计，总比没有这种性质的估计好些.

无偏估计的实例在前面已提到一些. 如例 1.1 和 1.2 中，用样本废品率估计整批废品率；例 1.4 中，分别用样本均值 $\bar{X}$ 和样本方差 S^2 估计正态分布 $N(a, \sigma^2)$ 中的 a 和 σ^2. 例 1.5 中用样本

均值估计总体分布均值 $\frac{1}{\lambda}$. 一般地，如(2.3)式指出的，用样本原点矩 a_{nk} 估计总体原点矩 α_k 是无偏的，但样本中心矩 m_{nk} 不是总体中心矩 μ_k 的无偏估计. 不过可以证明：偏差随样本大小 n 增加而变得可忽略不计，一般有 $Em_{nk}=\mu_k+O\left(\frac{1}{n}\right)$.

一个参数的无偏估计往往不止一个. 另外，也有些情况，其中根本没有无偏估计.

例 2.13 设样本 $X\sim B(n, p)$, n 已知而 p 为未知参数(样本大小为 1), $g(p)=\sin p$. 因为 X 只取 0, 1, …, n 这些值，为定义一个估计量 $\hat{g}$, 只须指出 $\hat{g}(i)$之值 a_i 即可，$i=0, \cdots, n$. 若 $\hat{g}$ 为无偏，则依定义 2.4，应有

$$E_p\hat{g}(X)=\sum_{i=0}^{n}a_i\binom{n}{i}p^i(1-p)^{n-i}=g(p)=\sin p,\ 0<p<1.$$

但 $\sum_{i=0}^{n}a_i\binom{n}{i}p^i(1-p)^{n-i}$ 是 p 的 n 阶多项式，它不可能在一个区间 (0, 1) 上处处等于一个超越函数 $\sin p$. 以此知 $\sin p$ 没有无偏估计. 其他例子见本章习题.

无偏估计既然一般不唯一，就存在一个从其中挑选的问题. 为此就必须引进一定的准则作为挑选的根据. 下面采用**均方误差准则**，这是统计上常用的一个准则. 我们设 $g(\theta)$ 为一维的.

设用 $\hat{g}(X)$估计 $g(\theta)$. 则一般 $\hat{g}(X)-g(\theta)$不为 0(如前指出的，即使 $\hat{g}$ 为无偏，也是如此)，它是估计量 $\hat{g}$ 在这一具体场合(具体的样本 X)下的误差. 把这误差平方以消除符号的影响，得 $[\hat{g}(X)-g(\theta)]^2$. 这个量仍是随机的，再计算其均值，以得到一个整体性指标 $E_\theta[\hat{g}(X)-g(\theta)]^2$, 这就是估计量 $\hat{g}$ 的**均方误差**，常记为 $\mathrm{MSE}_\theta(\hat{g})$(MSE 是 Mean Square Error 的缩写). 在这个准则下，若 $\hat{g}_1$ 和 $\hat{g}_2$ 为两个估计量而 $\mathrm{MSE}_\theta(\hat{g}_1)\leqslant\mathrm{MSE}_\theta(\hat{g}_2)$对一切 $\theta\in\Theta$, 且不等号至少对 Θ 中的一个 θ 值成立，则 $\hat{g}_1$ 优于 $\hat{g}_2$.

当 $\hat{g}$ 为 $g(\theta)$的无偏估计时，有 $\mathrm{MSE}_\theta(\hat{g})=\mathrm{Var}_\theta(\hat{g})$, 即 $\hat{g}$ 的方差. 于是按均方误差准则，在无偏估计类中，方差愈小愈优. 这个

考虑引导到下面的定义：

定义 2.5 设 $\hat{g}$ 为 $g(\theta)$ 的一个无偏估计. 若对 $g(\theta)$ 的任一无偏估计 $\hat{g}_1$，都有

$$\mathrm{Var}_\theta(\hat{g}) \leqslant \mathrm{Var}_\theta(\hat{g}_1), \text{ 对一切 } \theta \in \Theta. \tag{2.26}$$

则称 $\hat{g}$ 是 $g(\theta)$ 的一个**一致最小方差无偏估计**，简称为 UMVUE (是 Uniformly Minimum Variance Unbiased Estimate 的缩写)。

本节下面几段中将讨论求 UMVUE 的几个重要方法(确切地说，是证明某一特定估计为 UMVUE 的方法). 在这项工作中，下面的性质起简化问题的作用，有重要意义：

引理 2.1 设 $T=T(X)$ 是一个充分统计量，而 $\hat{g}(X)$ 是 $g(\theta)$ 的一个无偏估计. 则存在可表为 T 的函数的无偏估计 $h(T(X))$，使

$$\mathrm{Var}_\theta(h(T(X)) \leqslant \mathrm{Var}_\theta(\hat{g}(X)), \tag{2.27}$$

等号当且仅当 $\hat{g}(X)$ 能表为 $T(X)$ 的函数时才成立.

证 因 T 为充分统计量，故按定义，在给定 T 时，样本 X 的条件分布与参数 θ 无关，因此，在给定 T 之下，$\hat{g}(X)$ 的条件期望 $E_\theta(\hat{g}(X)|T)$ 也与 θ 无关，可记为 $h(T)$. 现因 $\hat{g}$ 为无偏估计，有

$$E_\theta h(T(X)) = E_\theta E_\theta(\hat{g}(X)|T) = E_\theta(\hat{g}(X)) = g(\theta),$$
$$\text{一切 } \theta \in \Theta.$$

因此，$h(T(X))$ 为 $g(\theta)$ 的无偏估计. 现任取 $\theta_0 \in \Theta$，有

$$h^2(T(X)) = [E_{\theta_0}(\hat{g}(X)|T)]^2 \leqslant E_{\theta_0}(\hat{g}^2(X)|T).$$

最后两项之差为 $\mathrm{Var}_{\theta_0}(\hat{g}(X)|T)$. 当且仅当在给定 T 的条件下 $\hat{g}(X)$ 为一常数，也就是 $\hat{g}(X)$ 可表为 T 的函数时，它才等于 0. 因此 $E_{\theta_0}h^2(T(X)) \leqslant E_{\theta_0}E_{\theta_0}(\hat{g}^2(X)|T) = E_{\theta_0}\hat{g}^2(X)$，而

$$\begin{aligned}
&\mathrm{Var}_{\theta_0}(h(T(X)) \\
&= E_{\theta_0}h^2(T(X)) - g^2(\theta_0) \leqslant E_{\theta_0}\hat{g}^2(X) - g^2(\theta_0) \\
&= \mathrm{Var}_{\theta_0}(\hat{g}(X)),
\end{aligned}$$

等号当且仅当 $\hat{g}(X)$ 可表为 T 的函数时才成立. 引理证毕.

据此，在有充分统计量 T 时，为求 UMVUE，只须考虑能表为

T 的函数的无偏估计类。

（二）零无偏估计法

以下我们沿用上一段的记号．本段介绍一个一般性定理，用以判明某一估计量是否为 UMVUE.

定理 2.1 设 $\hat{g}(X)$ 为 $g(\theta)$ 之一无偏估计，$\mathrm{Var}_\theta(\hat{g}(X))<\infty$ 对任何 $\theta\in\Theta$，且对任何满足条件"$E_\theta l(X)=0$ 对一切 $\theta\in\Theta$"的统计量 l，必有

$$\mathrm{Cov}_\theta(\hat{g}(X),\ l(X))=E_\theta(\hat{g}(X)l(X))=0,\quad \text{一切 } \theta\in\Theta, \tag{2.28}$$

则 $\hat{g}(X)$ 是 $g(\theta)$ 的 UMVUE.

从形式上看，条件"$E_\theta l(X)=0$ 对一切 $\theta\in\Theta$"可解释为"$l(X)$ 是零的无偏估计"．以此得到本方法的名称．

证 设 $\hat{g}_1(X)$ 为 $g(\theta)$ 的任一无偏估计．记 $l(X)=\hat{g}_1(X)-g(X)$，则 $l(X)$ 为零的无偏估计，故(2.28)成立，因而

$$\begin{aligned}\mathrm{Var}_\theta(\hat{g}_1(X))&=\mathrm{Var}_\theta(\hat{g}(X)+l(X))\\&=\mathrm{Var}_\theta(\hat{g}(X))+\mathrm{Var}_\theta(l(X))\\&\quad+2\,\mathrm{Cov}_\theta(\hat{g}(X),\ l(X))\\&=\mathrm{Var}_\theta(\hat{g}(X))+\mathrm{Var}_\theta(l(X))\\&\geqslant\mathrm{Var}_\theta(\hat{g}(X)).\end{aligned}$$

这证明了所要的结果．

从定理的内容看出，它是一个验证某个特定的估计量 $\hat{g}$ 为 UMVUE 的工具．至于这个特定的 $\hat{g}$ 从何而来，只能笼统地说，一般是基于直观的想法提出的、看来是很好的估计．如果我们无法用这种或那种考虑"捉"到这个特定的 $\hat{g}$，这个定理就无所帮助．条件(2.28)的验证也不容易，因为零的无偏估计很多．但是，前面一些例子中提到的几个常用估计，都可用此法验证其为 UMVUE.

例 2.14 设一事件 A 的概率 p 未知．独立地作 n 次试验，记 $X_i=1$ 或 0，视第 i 次试验中事件 A 发生与否而定，$i=1,\ \cdots,\ n$，求 p 的 UMVUE.

记 $T(X)=X_1+\cdots+X_n(X=(X_1, \cdots, X_n))$. 如在例 1.19 中证明的,$T$ 为充分统计量. 按引理 2.1,可以局限于可表为 T 的函数的无偏估计类. 现取 $\hat{g}=T/n$, 来验证它满足定理 2.1 的两个条件. 首先很明显, $\hat{g}$ 是 p 的无偏估计, 且 $\mathrm{Var}_p(\hat{g})<\infty$ 对 $0\leqslant p\leqslant 1$. 现设 $l=l(T)$ 为零的无偏估计, 并记 $a_i=l(i)$, $i=0, \cdots, n$, 则因 $T\sim B(n, p)$,有

$$E_p l=\sum_{i=0}^{n} a_i\binom{n}{i}p^i(1-p)^{n-i}=0,\ 0\leqslant p\leqslant 1.$$

约去因子 $(1-p)^n$, 并记 $s=p/(1-p)$ (s 取值于 $(0, \infty)$), 上式可改写为 $\sum_{i=1}^{n} a_i\binom{n}{i}s^i=0$, 一切 $s\in(0, \infty)$. 由此知 $a_i\binom{n}{i}=0$, $i=0, \cdots, n$, 即 $a_i=0$, $i=0, \cdots, n$. 因而估计量 l 只取 0 为值. 这时 (2.28) 当然成立, 由此证明了: 频率 T/n 确是 p 的 UMVUE.

例 2.15 回到例 1.5, 求总体分布均值 $\frac{1}{\lambda}$ 的 UMVUE.

在例 1.23 中已提出: $T(X)=X_1+\cdots+X_n$ 为充分统计量. 又在例 1.13 中, 已求得 T 的密度函数为 (1.20). 取 $\hat{g}=T/n$, 则 $\hat{g}$ 为 $\frac{1}{\lambda}$ 的无偏估计, 且 $\mathrm{Var}_\lambda(\hat{g})<\infty$ 对 $\lambda>0$. 现设 $l=l(T)$ 为零的无偏估计, 则由 (1.20) 式, 有

$$0=E_\lambda l(T)=\frac{1}{(n-1)!}\int_0^\infty \lambda^n e^{-\lambda x}x^{n-1}l(x)dx,\ \lambda>0,$$

即 $\int_0^\infty l(x)e^{-\lambda x}x^{n-1}dx=0$. 两边对 λ 求导数, 得 $\int_0^\infty x l(x)e^{-\lambda x}x^{n-1}dx=0$, 即 $E_\lambda(Tl(T))=0$ 对一切 $\lambda>0$, 因而 $\hat{g}$ 适合条件 (2.28). 这证明了样本均值 $\bar{X}=T/n$ 是 $\frac{1}{\lambda}$ 的 UMVUE.

例 2.16 回到例 1.24, 要求 θ 的 UMVUE. 在该例中已证明: 统计量 $T=T(X)=\max(X_1, \cdots, X_n)$ 是充分的. 易算出 T 的分布, 因为 $P_\theta(T<t)=P_\theta(X_1<t, \cdots, X_n<t)=\prod_{i=1}^{n}P_\theta(X_i<t)=t^n/\theta^n$ 当 $0<t<\theta$, 故 T 的概率密度为 nt^{n-1}/θ^n (当 $0<t<\theta$, 其他

处为 0). 取 $\hat{g}=\hat{g}(T)=\frac{n+1}{n}T$, 则

$$E_\theta\hat{g}=\int_0^\theta \theta^{-n}nt^{n-1}\frac{n+1}{n}tdt=\theta,\ \text{一切}\ \theta>0,$$

即 $\hat{g}$ 为 θ 的无偏估计. 现设 $l=l(T)$ 为零的无偏估计, 则应有 $\int_0^\theta l(x)nx^{n-1}/\theta^n dx=0$ 即 $\int_0^\theta l(x)x^{n-1}dx=0$ 对一切 $\theta>0$. 对 θ 求导, 得 $l(\theta)\theta^{n-1}=0$ 对一切 $\theta>0$, 因而 $l(\theta)=0$ 当 $\theta>0$. 即 l 只取 0 为值, 因而条件(2.28)成立. 这证明了: $\frac{n+1}{n}T$ 是 θ 的 UMVUE.

例 2.17 设 $X_1, \cdots, X_n \sim N(a,\sigma^2)$, 求 a 和 σ^2 的 UMVUE. 记 $T=(T_1, T_2)$, 其中 $T_1=\overline{X}$, $T_2=(n-1)S^2=\sum_{i=1}^n(X_i-\overline{X})^2$. 在例 1.21 中已证明了 T 是充分统计量. 又根据定理 1.1, 得到 T 的密度函数为

$$f_\theta(t_1, t_2)=\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}e^{-n(t_1-a)^2/2\sigma^2}$$

$$\times\left(2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)\sigma^{n-1}\right)^{-1}t_2^{(n-3)/2}e^{-t_2/2\sigma^2}. \quad (2.29)$$

(当 $-\infty<t_1<\infty$, $t_2>0$. 其他处为 0. 又 $\theta=(a, \sigma)$.)

先考虑 a 的估计. 令 $\hat{g}=\hat{g}(T)=T_1$, 则 $\hat{g}$ 为 $g(\theta)=a$ 的无偏估计, 且 $\mathrm{Var}_\theta(\hat{g})<\infty$. 现设 $l=l(T)$ 为零的无偏估计, 则 $E_\theta l(T)=0$. 按(2.29), 此可写为

$$\int_0^\infty\int_{-\infty}^\infty l(t_1, t_2)\exp\left(-\frac{1}{2\sigma^2}(n(t_1-a)^2+t_2)\right)dt_1dt_2=0, \quad (2.30)$$

对一切 $a\in(-\infty, \infty)$ 和 $\sigma>0$. 两边对 a 求导数, 得

$$\int_0^\infty\int_{-\infty}^\infty l(t_1, t_2)(t_1-a)\exp\left(-\frac{1}{2\sigma^2}(n(t_1-a)^2+t_2)\right)dt_1dt_2=0.$$

按(2.29), 此即

$$\mathrm{Cov}_\theta(l, \hat{g})=E_\theta[l(T)(T_1-a)]=0,\quad -\infty<a<\infty,\ \sigma>0,$$

因此条件(2.28)满足．这证明了 $\overline{X}$ 就是 a 的 UMVUE．完全同样的方法证明：$T_2/(n-1)=S^2$ 是 σ^2 的 UMVUE．

比较例 2.16 和例 2.17，我们看到这样一个现象：在例 2.16 中，总体均值为 $\theta/2$，其 UMVUE 是 $\frac{n+1}{2n}\max(X_1, \cdots, X_n)$，而在例 2.17 中，总体均值 a 的 UMVUE 为样本均值 $\overline{X}$．在例 2.16 中，$\overline{X}$ 比 $\frac{n+1}{2n}\max(X_1, \cdots, X_n)$ 要差些，而在例 2.17 中则反过来．因此，某一统计量，或者说，使用该统计量的统计推断方法，其优越性如何，不仅取决于这统计量本身，还要看统计模型是怎样的．

（三）充分-完备统计量法

本方法实际上可以看成是定理 2.1 的一个特例，只因它涉及一个重要的概念——**统计量的完备性**，我们把它摆在与定理 2.1 平行的地位．

设 T 为一充分统计量．据引理 2.1，若 $g(\theta)$ 的 UMVUE 存在，它必能表为 T 的函数．如果进一步知道：能表为 T 的函数的，$g(\theta)$ 的无偏估计唯一，则它就是 $g(\theta)$ 的 UMVUE．显然，当且仅当下述条件成立时，能表为 T 的函数的，$g(\theta)$ 的无偏估计才唯一："任何满足条件 $E_\theta l(T)=0$（一切 $\theta\in\Theta$）的 $l(T)$ 必为 0"．这个性质就取作为统计量 T 的完备性的定义：

定义 2.6 设 T 为一统计量（不必是充分的）．若对任何满足条件

$$E_\theta l(T(X))=0, \text{ 一切 } \theta\in\Theta \tag{2.31}$$

的 $l(T)$，都有

$$P_\theta(l(T(X))=0)=1, \text{ 一切 } \theta\in\Theta, \tag{2.32}$$

则称 T 是一个**完备统计量**．

注意，T 的完备性不仅取决于 T 的形状，还取决于样本 X 的分布族．完备性这个名称，来源于正交函数理论中的一类似概念．

为简单计，设统计量 $T(X)$ 有概率密度 $h_\theta(t)$，则(2.31)式可写为

$$\int l(t)h_\theta(t)dt=0, \text{一切 } \theta\in\Theta. \tag{2.33}$$

$\int l(t)h_\theta(t)dt=0$ 形式上可看成是“l 与 h_θ 正交”. 于是，条件(2.31)可说成是“l 与函数系 $\{h_\theta,\ \theta\in\Theta\}$ 正交”. 在正交函数论中，若 M 为一正交函数系，且不存在与 M 正交的非零函数，则称 M 为完备正交系. 由(2.33)看出，我们这里的完备性正好与此相当，不过我们不称密度函数系 $\{h_\theta,\ \theta\in\Theta\}$ 完备而称统计量 T 完备. 由于 $\{h_\theta,\ \theta\in\Theta\}$ 是由 T 决定的，这种称呼并不影响实质.

由完备统计量的定义并根据引理 2.1，或直接用定理 2.1，得到

定理 2.2 设 T 为一个完备充分统计量，$\hat{g}(T(X))$ 为 $g(\theta)$ 之一无偏估计，满足 $\mathrm{Var}_\theta(\hat{g}(T(X)))<\infty$ 对一切 $\theta\in\Theta$，则 $\hat{g}(T(X))$ 是 $g(\theta)$ 的唯一的 UMVUE(唯一性是在这样的意义下：若 $\hat{g}$ 和 $\hat{g}_1$ 都是 $g(\theta)$ 的 UMVUE，则 $P_\theta(\hat{g}\neq\hat{g}_1)=0$，对一切 $\theta\in\Theta$).

完备性概念是 Lehmann 和 Scheffe 在 1950 年提出来的. 定理 2.2 也是他们所证明，故常冠以他们的名称. 现在举几个例子.

例 2.18 在例 2.14 和例 2.16 中，我们都证明了在该处引进的充分统计量 T 满足完备性条件，由此可知，在这两个例子中得出的 UMVUE 都是唯一的.

例 2.19 回到例 1.1. 考虑批废品率 $\dfrac{M}{N}$ 的估计(此处 N 已知，M 为参数). 令 $T(X)=X_1+\cdots+X_n$. 在例 1.19 中证明了 T 为充分统计量，又 T/n 为 M/N 的无偏估计. 因此据定理 2.2，只要证明了 T 为完备统计量，就可以肯定 T/n 即 $\bar{X}$ 是 $\dfrac{M}{N}$ 的唯一的 UMVUE. 设 $l(T)$ 为一统计量，满足条件

$$E_M l(T)=\sum_{i=0}^{M} l(i)\binom{M}{i}\binom{N-M}{n-i}\Big/\binom{N}{n}=0,\ M=0,1,\cdots,N.$$

令 $M=0$，由上式得 $l(0)=0$. 再令 $M=1$，得

$$\binom{N-M}{n}l(0)+\binom{M}{1}\binom{N-M}{n-1}l(1)=0,$$

故 $l(1)=0$. 继续下去，可得 $l(i)=0$，对 $i=0, 1, \cdots, n$. 这证明了 T 的完备性.

例 2.20 回到例 2.15. 我们来证明，此例中引进的统计量 $T=X_1+\cdots+X_n$ 是完备的. 这样就证明了：在该例中得出的估计 T/n 即 $\bar{X}$，是 $\frac{1}{\lambda}$ 的唯一的 UMVUE.

事实上，若 $E_\lambda l(T)=0$ 对一切 $\lambda>0$，则有

$$\int_0^\infty l(x)x^{n-1}e^{-\lambda x}dx=0,\ \lambda>0.$$

这就是说，函数 $l(x)x^{n-1}$ 的 Laplace 变换为 0. 按 Laplace 变换的唯一性，即得 $l(x)x^{n-1}=0$，因而 $l(x)=0$. 明所欲证.

对重要的正态分布参数(例(2.17))也可以证明同样的结果. 现在我们引进一个判别完备性的一般定理，它包括了一些常见的情况. 定理的证明在此不给出，可参看陈希孺《数理统计引论》p. 80～82.

定理 2.3 设样本 X 有概率函数

$$f(x,\ \theta)=C(\theta)\exp(T_1(x)Q_1(\theta)+\cdots+T_k(x)Q_k(\theta))h(x),\ \theta\in\Theta. \tag{2.34}$$

这里 $C(\theta)$, $Q_1(\theta)$, $\cdots$, $Q_k(\theta)$ 只与 θ 有关，$T_1(x)$, $\cdots$, $T_k(x)$ 和 $h(x)$ 只与样本 x 有关，则统计量

$$T(X)=(T_1(X),\ \cdots,\ T_k(X))$$

是充分统计量. 又若 $\{(Q_1(\theta),\ \cdots,\ Q_k(\theta):\theta\in\Theta\}$ 作为 R^k 中的子集有内点，则 T 也是完备的(注意：T 的充分性是定理 1.5 的直接推论).

形如(2.34)的分布族称为**指数型分布族**，它包含了许多在统计上重要的分布族.

例 2.21 设 $X_1, \cdots, X_n\sim N(a,\ \sigma^2)$，$\theta=(a,\ \sigma)$. 易见 X 的概率密度可写为(2.34)的形状，其中 $k=2$，$T_1(x)=\sum_{i=1}^n x_i$，$T_2(x)$

$=\sum_{i=1}^{n} x_i^2$, $Q_1(\theta)=a/\sigma^2, Q_2(\theta)=-1/2\sigma^2, C(\theta)=(\sqrt{2\pi}\sigma)^{-n}\cdot\exp\left(-\frac{na^2}{2\sigma^2}\right)$, $h\equiv 1$. 集合 $\left\{\left(\frac{a}{\sigma^2}, -\frac{1}{2\sigma^2}\right): -\infty<a<\infty, \sigma>0\right\}$是 R^2 中的半平面，当然有内点. 由此依定理 2.3，知 $T=(T_1, T_2)$为充分完全统计量. 而作为 T 的函数的估计量

$$\overline{X}=T_1/n,\ S^2=(T_2-T_1^2/n)/(n-1)$$

分别是 a 和 σ^2 的无偏估计. 故按定理 2.2，知 $\overline{X}$ 和 S^2 分别是 a 和 σ^2的唯一的 UMVUE.

(四)C-R 不等式法

此法的思想如下：考虑 $g(\theta)$的一切无偏估计的类 U_g. U_g 中的估计的方差有下确界 $d_g(\theta)=\inf\{\mathrm{Var}_\theta(\hat{g}):\hat{g}\in U_g\}$. 若能找到 $g(\theta)$的一个无偏估计 g^*，使 $\mathrm{Var}_\theta(g^*)=d_g(\theta)$对一切 $\theta\in\Theta$，则 g^* 显然就是 $g(\theta)$的 UMVUE. 在 1945 年和 1946 年，C. R. Rao 和 H. Cramer 独立地得到了 $d_g(\theta)$的一个下方估计，即找到了一个表达式 $d_g^*(\theta)$，使 $d_g^*(\theta)\leqslant d_g(\theta)$. 所得到的不等式称为 **Cramer-Rao 不等式**，简称 **C-R 不等式**.

这个不等式的严格推导要假定样本分布满足一系列的正则性条件. 下面我们先形式地把它推出来，再指出推导过程中所用到的假定.

设样本 X 有概率函数 $f(x, \theta)$，且为确定计，设 $f(x, \theta)$是概率密度(离散的情况可以同样处理). 参数 θ 是一维的，在 R' 的一个开区间 $\Theta=(a, b)$上取值(a 可为 $-\infty$，b 可为 $+\infty$)，$g(\theta)$为待估函数. 设 $\hat{g}(x)$为 $g(\theta)$之一无偏估计，则依定义，有

$$\int \hat{g}(x)f(x, \theta)dx=g(\theta),\ \theta\in\Theta. \tag{2.35}$$

积分的范围是样本空间$\mathscr{X}$. 两边对 θ 求导，得

$$\int \hat{g}(x)\frac{\partial f(x, \theta)}{\partial\theta}dx=g'(\theta). \tag{2.36}$$

又注意到

$$\int f(x, \theta)dx=1, \tag{2.37}$$

两边对 θ 求导，得

$$\int \frac{\partial f(x, \theta)}{\partial \theta} dx=0. \tag{2.38}$$

由(2.36)和(2.38)，得

$$\int[\hat{g}(x)-g(\theta)] \frac{\partial f(x, \theta)}{\partial \theta} dx=g'(\theta).$$

将此式写为

$$\int\{(\hat{g}(x)-g(\theta))\sqrt{f(x, \theta)}\}\left\{\frac{1}{\sqrt{f(x, \theta)}} \frac{\partial f(x, \theta)}{\partial \theta}\right\}dx$$
$$=g'(\theta).$$

用 Cauchy-Schwarz 不等式，得

$$\int[\hat{g}(x)-g(\theta)]^2 f(x, \theta)dx \cdot \int\left(\frac{1}{f(x, \theta)} \frac{\partial f(x, \theta)}{\partial \theta}\right)^2 f(x, \theta)dx$$
$$\geqslant[g'(\theta)]^2. \tag{2.39}$$

上式左边第一项是 $\mathrm{Var}_\theta(\hat{g}(X))$，第二项是 $E_\theta\left(\frac{\partial \log f(X, \theta)}{\partial \theta}\right)^2$.

于是得到

$$\mathrm{Var}_\theta(\hat{g}(X))\geqslant[g'(\theta)]^2/E_\theta\left(\frac{\partial \log f(X, \theta)}{\partial \theta}\right)^2. \tag{2.40}$$

特别，当 $g(\theta)=\theta$ 时，有

$$\mathrm{Var}_\theta(\hat{g}(X))\geqslant 1/E_\theta\left(\frac{\partial \log f(X, \theta)}{\partial \theta}\right)^2. \tag{2.41}$$

不等式(2.40)，或其特例(2.41)，就是 C-R 不等式. 它指出了 $g(\theta)$ 的无偏估计 $\hat{g}$ 的方差不可逾越的一个下界，即(2.40)右边.它就是我们上文提到的 $d_g^*(\theta)$.

特别，在一个重要的情况，其中 $X=(X_1, \cdots, X_n)$，而 $X_1, \cdots, X_n$ 独立同分布，总体概率函数为 f_θ. 这时，$f(x, \theta)=f_\theta(x_1)\cdots f_\theta(x_n)$，而

$$\frac{\partial \log f(x, \theta)}{\partial \theta}=\sum_{i=1}^n \frac{\partial \log f_\theta(x_i)}{\partial \theta}. \tag{2.42}$$

当 $i\neq j$ 时，因 X_i 和 X_j 独立，有

$$E_\theta\left(\frac{\partial\log f_\theta(X_i)}{\partial\theta}\ \frac{\partial\log f_\theta(X_j)}{\partial\theta}\right)$$

$$=E_\theta\left(\frac{\partial\log f_\theta(X_i)}{\partial\theta}\right)E_\theta\left(\frac{\partial\log f_\theta(X_j)}{\partial\theta}\right). \tag{2.43}$$

但
$$E_\theta\left(\frac{\partial\log f_\theta(X_i)}{\partial\theta}\right)=\int\frac{1}{f_\theta(x_i)}\frac{\partial f_\theta(x_i)}{\partial x_i}f_\theta(x_i)dx_i$$

$$=\int\frac{\partial f_\theta(x_i)}{\partial\theta}\ dx_i=\frac{\partial}{\partial\theta}\int f_\theta(x_i)\,dx_i=\frac{\partial}{\partial\theta}(1)=0,$$

故(2.43)式为0. 把(2.42)式两边平方再求均值，得

$$E_\theta\left(\frac{\partial\log f(X,\theta)}{\partial\theta}\right)^2=\sum_{i=1}^n E_\theta\left(\frac{\partial\log f_\theta(X_i)}{\partial\theta}\right)^2$$

$$=nE_\theta\left(\frac{\partial\log f_\theta(X_1)}{\partial\theta}\right)^2. \tag{2.44}$$

于是得到这一情况下的 C-R 不等式(取(2.41)的特例)：

$$\mathrm{Var}_\theta(\hat{g}(X))\geqslant\frac{1}{nI(\theta)}, \tag{2.45}$$

其中

$$I(\theta)=E_\theta\left(\frac{\partial\log f_\theta(X_1)}{\partial\theta}\right)^2=\int\frac{1}{f_\theta(t)}\left(\frac{\partial f_\theta(t)}{\partial t}\right)^2dt. \tag{2.46}$$

现在回过头来看看，在导出(2.40)式的过程中，用到了那些条件. 首先不待言，$g'(\theta)$必须存在. 其次，(2.35)式左边必须能在积分号下对 θ 求导，才能推出(2.36). 最后，(2.37)式左边也须在积分号下对 θ 求导以得到(2.38). 这些条件麻烦之处，在于它不仅与样本分布 $f(x,\theta)$有关，还与所讨论的无偏估计 $\hat{g}$ 有关. 可以建立某种较易验证的充分条件，下面就是其一：

定理 2.4 设 $X_1, \cdots, X_n$ 为简单随机样本，总体有概率函数 f_θ. 参数 θ 属于一开区间 $\Theta=(a,b)$，$g(\theta)$为在(a,b)上可微的待估函数. 设存在函数 $G(t,\theta)$，满足以下的条件：

1° $E_\theta G^2(X_1,\theta)<\infty$，对一切 $\theta\in\Theta$.

2° 对任何 $\theta\in\Theta$，存在 $\varepsilon_\theta>0$，使当$|\psi-\theta|<\varepsilon_\theta$时有

$$\left|\frac{\partial f_\psi(t)}{\partial\psi}\Big/f_\psi(t)\right|\leqslant G(t,\theta),$$

则当 $\hat{g}(X)$为 $g(\theta)$之一无偏估计时，必有(2.40).

本定理的证明是纯分析的，就是验证在这两个条件下，(2.35)和(2.37)在积分号下对 θ 求导为合法．仔细证明过程在此从略，可参看陈希孺《数理统计引论》的引理 2.2.1，本定理是其一特例．

下面举几个例子说明用 C-R 不等式求 UMVUE．在这些例子中都不难验证定理 2.4 的条件 1°、2°成立．

例 2.22 回到例 2.14．在本例中，总体有概率函数 $f_p(1)=p$，$f_p(0)=1-p$，可写为 $f_p(t)=p^t(1-p)^{1-t}$，$t=0, 1$．$\partial\log f_p(X_1)/\partial p=\dfrac{X_1-p}{p(1-p)}$．于是 $I(p)=\dfrac{1}{p(1-p)}$．由(2.45)，知 p 的任一无偏估计的方差不小于 $p(1-p)/n$．但无偏估计 $\bar{X}$ 的方差达到这个界限，因此就是 p 的 UMVUE．

例 2.23 回到例 2.2．总体分布为 Poisson 分布(2.7)在本例有 $\partial\log f_\theta(X_1)/\partial\theta=(X_1-\theta)/\theta$，由此得 $I(\theta)=\dfrac{1}{\theta}$．当样本大小为 n 时，θ 的无偏估计方差的下降为 θ/n．无偏估计 $\bar{X}$ 的方差达到这个下限，因此就是 θ 的 UMVUE．

例 2.24 回到例 2.15．总体密度为 $\lambda e^{-\lambda x}$，得

$$\frac{\partial\log f_\lambda(X_1)}{\partial\lambda}=(1-\lambda X_1)/\lambda,$$

有 $I(\lambda)=\dfrac{1}{\lambda}$．待估函数为 $g(\lambda)=\dfrac{1}{\lambda^2}$，得 C-R 下界为 $\left(\dfrac{1}{\lambda^2}\right)^2\Big/ n\dfrac{1}{\lambda^2}=\dfrac{1}{n\lambda^2}$．无偏估计 $\bar{X}$ 的方差正好达到这个下界，故就是 $\dfrac{1}{\lambda}$ 的 UMVUE．

以上除例 2.23 外，都是在前面已证明过的．例 2.23 也很容易用定理 2.1 或 2.2 解决．甚至用前面的解法更彻底，因为依定理 2.2，可知以上各例求出的 UMVUE 都唯一，而在此处则看不出唯一性．这一事实其实是普遍的．可以证明：凡是能用 C-R 不等式法求得 UMVUE 的场合，必能用定理 2.2(或定理 2.1)求得．从这一点看，C-R 不等式作为一种求 UMVUE 的方法，意义是不大的．但我们要注意两点：一是 C-R 不等式法，较基于定理 2.2

的方法为早．二是C–R不等式在数理统计理论上还有些其他的用处．以下要谈到的两个问题可以归入这个范畴，不过，在本教程中，我们没有机会去考察这个不等式的多方面应用．

1．估计的效率和有效估计　为简单计就考察θ的估计．据(2.45)，θ的任一无偏估计$\hat{\theta}$的方差$\mathrm{Var}_\theta(\hat{\theta})$都不小于$\frac{1}{nI(\theta)}$．比值

$$e_{\hat{\theta}}(\theta)=\frac{1}{nI(\theta)}\Big/\mathrm{Var}_\theta(\hat{\theta})=[nI(\theta)\mathrm{Var}_\theta(\hat{\theta})]^{-1} \qquad (2.47)$$

可以作为$\hat{\theta}$的方差与下界$\frac{1}{nI(\theta)}$的差距的一种衡量，称为无偏估计$\hat{\theta}$的**效率**．一般有$e_{\hat{\theta}}(\theta)\leqslant 1$，等号(对一切$\theta$)当且仅当$\hat{\theta}$为UMVUE时成立．当$e_{\hat{\theta}}(\theta)=1$(一切$\theta$)时，称$\hat{\theta}$为**有效估计**．在所述的意义下，有效估计与UMVUE是一回事．

这个概念有如下的缺点：在一些情况下，没有任何无偏估计能达到C–R不等式规定的下界，因此，在这种情况下不存在有效估计．另外，C–R不等式的成立有一定的条件．当这些条件不成立时，C–R不等式可以不对．这时，依据它所提供的下限去定义效率就不合理了．看下面的例子．

例2.25　考察例2.16．此处总体密度为$f_\theta(t)=\frac{1}{\theta}$当$0<t<\theta$，他处为0．注意对固定的$t>0$，$f_\theta(t)$作为$\theta$的函数，在$\theta=t$点不连续，当然就不存在偏导数．若形式地按公式计算，则得$\partial\log f_\theta(X_1)/\partial\theta=-\frac{1}{\theta}$，而$I(\theta)=1/\theta^2$．C–R不等式提供的下界为$\theta^2/n$．但若取例2.16中的无偏估计$\frac{n+1}{n}T=\frac{n+1}{n}\max(X_1,\cdots,X_n)$，则根据例2.16中求得的$T$的概率密度，易算出$\mathrm{Var}_\theta\left(\frac{n+1}{n}T\right)=\frac{\theta^2}{n(n+2)}$．这比C–R不等式提供的下限$\theta^2/n$要小．

鉴于这种情况，有的学者把效率的定义作适当的修改．不过，这一概念总的说来在理论和实际上的意义不大，故我们不多去讨

论它了.

2. Fisher 信息量　设总体分布有概率函数 f_θ, 则由(2.46)式所定义的 $I(\theta)$, 称为该分布族的 **Fisher 信息量**.

这个概念可以作如下的解释. 我们前已指出: 在估计为无偏时, 其方差愈小愈好. 直观上这表示估计值有更多地机会出现在 θ 的附近, 而导致较小的误差. 为说明方便, 暂时认为 C-R 不等式所提供的下限 $\frac{1}{nI(\theta)}$ 可以达到. 这时, $nI(\theta)$ 愈大, 则表示参数 θ 可以估得愈精, $nI(\theta)$ 与 n 和 $I(\theta)$ 都成比例. n 是样本大小, 这意味着若以估计量方差的倒数作为估计量精度的指标, 则精度与 n 成正比. 比例因子, 即 $I(\theta)$, 反映总体分布的一种特性. 就是说, 某总体分布的 $I(\theta)$ 愈大, 意味着该总体的参数愈容易估计, 或者说, 该总体模型本身提供的信息量愈多. 故此有理由把 $I(\theta)$ 视为一种衡量总体模型所含信息的量——信息量. $I(\theta)$ 也可以解释成单个样本提供的信息量. 由于样本 $X_1, \cdots, X_n$ 独立同分布, 它们的地位是平等的, 故每个样本应当提供同样多的信息, 即整个样本 $(X_1, \cdots, X_n)$ 所含信息量为 nc. c 的值是反映模型"易于认识"的程度. 此处取之为 $c=I(\theta)$ 是从估计量方差的角度着眼的.

可以看这样一个例子. 为估计物重 a, 甲、乙两人分别用两把天平各秤 n 次. 甲的天平较精, 其每次秤量值的分布为 $N(a, 1)$. 乙的天平较差, 其每次秤量值的分布为 $N(a, 2^2)$. 按公式(2.46), 甲、乙二人每次量测所提供的信息量分别为 $1/1^2=1$ 和 $1/2^2=1/4$. 在此, 结果是自然而合理的: 甲的天平较精, 他秤出的结果自应包含关于物重 a 的更多的信息量, 或者说, 在甲的情况下 a 更容易估得精确些.

Fisher 信息量 $I(\theta)$ 的重要意义在于, 在点估计的大样本理论的研究中, 它要起相当的作用. 下一节中我们会有机会看到这一点, 正是 Fisher 在二十年代关于点估计理论的研究中定义了这个量, 故后人称之为 Fisher 信息量. 这个量与 C-R 不等式发生关系, 也并不能算是一种巧合.

以上所讨论的都是参数 θ 为一维的情况，对高维的情况也可以建立类似的结果。为此先引进一个记号。设 $A=(a_{ij})$ 和 $B=(b_{ij})$ 是同阶的非负定方阵。若 $A-B$ 是非负定的，则记为 $A\geqslant B$，这时必有 $a_{ii}\geqslant b_{ii}$，对一切 i。

现设 $\theta=(\theta_1, \cdots, \theta_k)$，总体概率函数仍记为 f_θ，$X_1, \cdots, X_n$ 为自此总体中抽得的简单随机样本。设 $\hat{\theta}=\hat{\theta}(X_1, \cdots, X_n)=(\hat{\theta}_1, \cdots, \hat{\theta}_k)$ 是 θ 的一个无偏估计。以 $\mathrm{Cov}_\theta(\hat{\theta})$ 记 $\hat{\theta}$ 的协方差阵，它是一个 k 阶非负定方阵，其 (i, j) 元为 $E_\theta[(\hat{\theta}_i-\theta_i)(\hat{\theta}_j-\theta_j)]$，则在类似于 θ 为一维时的条件之下，可以证明

$$\mathrm{Cov}_\theta(\hat{\theta})\geqslant(nI(\theta))^{-1}. \tag{2.48}$$

这里 $I(\theta)=(I_{ij}(\theta))$ 是一个 k 阶正定方阵，且

$$I_{ij}(\theta)=E_\theta\left[\frac{\partial\log f_\theta(X_1)}{\partial\theta_i}\ \frac{\partial\log f_\theta(X_1)}{\partial\theta_j}\right],\ i, j=1, \cdots, k, \tag{2.49}$$

这就是多维的 C-R 不等式。由(2.48)知，若以 $(I^*_{ij}(\theta))$ 记 $I(\theta)$ 的逆矩阵 $(I(\theta))^{-1}$，则有

$$\mathrm{Var}_\theta(\hat{\theta}_i)\geqslant I^*_{ii}(\theta)/n,\ i=1, \cdots, k. \tag{2.50}$$

这给出了每个分量 θ_i 的无偏估计方差的下限。

例 2.26 设 $X_1, \cdots, X_n\sim N(a, \sigma^2)$，$\theta=(a, \sigma^2)$（即 $\theta_1=a$，$\theta_2=\sigma^2$）。此处 $f_\theta(t)=(2\pi\theta_2)^{-1/2}e^{-(t-\theta_1)^2/2\theta_2}$。由此知

$$\frac{\partial\log f_\theta(X_1)}{\partial\theta_1}=\frac{X_1-\theta_1}{\theta_2},\quad \frac{\partial\log f_\theta(X_1)}{\partial\theta_2}=\frac{\theta_2-(X_1-\theta_1)^2}{2\theta_2^2}.$$

由此易算出 $I_{11}(\theta)=1/\sigma^2$，$I_{22}(\theta)=\dfrac{1}{2\sigma^4}$，$I_{12}(\theta)=I_{21}(\theta)=0$。而 $(I(\theta))^{-1}$ 的各元为 $I^*_{11}(\theta)=\sigma^2$，$I^*_{22}(\theta)=2\sigma^4$，$I^*_{12}(\theta)=I^*_{21}(\theta)=0$。因此按(2.50)式知，$a$ 和 σ^2 的任何无偏估计，其方差分别不能超过 σ^2/n 和 $2\sigma^4/n$。前一个下限可以达到，就是 $\bar{X}$，而后一个界限无法达到。事实上，在例 2.21 中我们已证明了 S^2 是 σ^2 的 UMVUE。利用 $(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$ 以及 $\mathrm{Var}(\chi^2_{n-1})=2(n-1)$ 的事实，知 $\mathrm{Var}_\theta(S^2)=\dfrac{2\sigma^4}{n-1}>\dfrac{2\sigma^4}{n}$。因为 S^2 是 UMVUE，不存在

σ^2 的无偏估计，其方差能达到下限 $2\sigma^4/n$.

由本例看出，即使在估计正态总体方差这样简单的场合，C-R 不等式的下界也不能达到. 因此以这个不等式作为找 UMVUE 的方法是不理想的. 也有一些学者研究改进这个不等式的问题.

出现在(2.48)式中的、由(2.49)定义的方阵$I(\theta)$,称为**Fisher 信息阵**.

§2.3 点估计的大样本理论

关于“大样本”、“小样本”的含义，已在§1.4(四)中解释过了. 按那里所阐述的观点，点估计的大样本理论的对象，就是研究当样本大小 $n\to\infty$ 时，点估计量的极限性质（大样本性质）. 固然，在任何具体应用中，样本大小 n 都是有限的. 但是，如在§1.4(四)中所说明的，大样本性质对于衡量一个统计推断方法(在此为点估计)的优劣也有重要意义，并提供了一种近似地计算其性能指标的方法. 对这后一点我们有必要指出：目前大样本理论的成果对这一目的而言还是远不能令人满意的. 因为它未能给出当样本大小 n 固定时，种种有关的量的值与其极限值的差别的有用估计.

点估计的大样本理论在战后几十年来发展很快. 著作文献可以说是汗牛充栋. 这里只能介绍几个最基本的概念和结果.

(一)相合估计

以 $X^{(n)}$记样本($X_1, \cdots, X_n$)，n 是样本大小. 我们这里没有假定 $X_1, \cdots, X_n$ 独立同分布，虽则这是最常见的情况. 设 $X^{(n)}$ 的分布(样本分布)依赖于参数 θ，参数空间记为 Θ，$g(\theta)$为定义在 Θ 上的已知函数，而 $\hat{g}(X^{(n)})$为 $g(\theta)$的点估计量. 如果当样本大小 n 无限增加时，估计值 $\hat{g}(X^{(n)})$在某种意义下收敛于被估计的值 $g(\theta)$，就称 $\hat{g}(X^{(n)})$为 $g(\theta)$在相应意义下的**相合估计**. “相合”一词可形象地理解为 $\hat{g}(X^{(n)})$“合”于 $g(\theta)$.

相合性可以说是对估计量的一个起码而合理的要求. 试想：

若不论作多少次试验，也不能把$g(\theta)$估计到任意指定的精确程度，则这个估计量是否合用是值得怀疑的.

相合性的具体意义取决于收敛的意义. 下面给出正式的定义.

定义 2.7 如果当样本大小 $n\to\infty$ 时，$\hat{g}(X^{(n)})$：1° 依概率收敛于 $g(\theta)$，即对任何 $\theta\in\Theta$ 及 $\varepsilon>0$ 有

$$\lim_{n\to\infty} P_\theta(|\hat{g}(X^{(n)})-g(\theta)|\geqslant\varepsilon)=0,$$

则称它为 $g(\theta)$ 的**弱相合估计**. 2° 以概率 1 收敛于 $g(\theta)$，即 $P_\theta(\lim_{n\to\infty}\hat{g}(X^{(n)})=g(\theta))=1$ 对任何 $\theta\in\Theta$，则称它为 $g(\theta)$ 的**强相合估计**. 3° r 阶矩收敛于 $g(\theta)$ $(r>0)$，即对任何 $\theta\in\Theta$ 有

$$\lim_{n\to\infty} E_\theta|\hat{g}(X^{(n)})-g(\theta)|^r=0,$$

则称它是 $g(\theta)$ 的 **r 阶矩相合估计**. 当 $r=2$ 时常称为**均方相合估计**.

根据概率论关于这几种收敛性的性质，一般地有：强相合$\Rightarrow$弱相合；对任何 $r>0$ 有：r 阶矩相合$\Rightarrow$弱相合. 反过来一般不成立. 又强相合与 r 阶矩相合之间没有包含关系.

现在来证明下面的一般性定理：

定理 2.5 设 $X_1, \cdots, X_n$ 为简单随机样本，$g(\theta)$有(2.5)的形状，G 为其变元的连续函数，而 $\hat{g}(X^{(n)})$是矩估计(2.6). 则 $\hat{g}(X^{(n)})$是 $g(\theta)$ 的强相合估计.

证 首先注意下面的简单事实：设 $f(y_1,\cdots,y_k)$在$(c_1, \cdots, c_k)$点连续，Y_{ni}，$i=1, \cdots, k$，$n=1, 2, \cdots$，都是随机变量，满足条件：$P(\lim_{n\to\infty} Y_{ni}=c_i)=1$，$i=1, \cdots, k$，则 $P(\lim_{n\to\infty} f(Y_{n1}, \cdots, Y_{nk})=f(c_1, \cdots, c_k))=1$. 根据 Kolmogorov 强大数律，有 $P(\lim_{n\to\infty} a_{ni}=\alpha_i)=1$，$i=1, \cdots, k$. 因为 $\mu_k=\sum_{r=0}^{k}(-1)^{k-r}\alpha_r\alpha_1^{k-r}$. 由(2.2)式及上面指出的事实，有 $P(\lim_{n\to\infty} m_{nk}=\mu_k)=1$，$k=1, 2, \cdots, l$，再利用证明开始处指出的事实，以及 G 的连续性，即得到所要的结果.

由这个定理可得出一些常见估计量的相合性. 例如，正态总

体 $N(a, \sigma^2)$ 中用样本均值 $\overline{X}$ 和样本方差 S^2 估计 a 和 σ^2, 是强相合估计, 也不难证明, S^2 是 σ^2 的均方相合估计(习题 15). 其实对任何 $r>0$, S^2 是 σ^2 的 r 阶矩相合估计). 例 2.5 中定义的偏度、峰度和变异系数, 其矩估计都是强相合估计.

另一类重要的估计是极大似然估计. 关于极大似然估计的相合性问题, 引起了许多统计学者的兴趣, 直到现在都不能说已经彻底解决了. 1946 年 Cramer 在一些条件下, 证明了似然方程有一个根是参数 θ 的弱相合估计. 由于似然方程的根不一定就是极大似然估计, 这个结果还没有解决极大似然估计的相合性问题. 直到 1949 年, Wald 才首次证明了极大似然估计的强相合性, 但所要求的条件很复杂. 嗣后一些学者继续进行研究, 希望在较少的条件下得到证明. 这些结果我们都不可能在此细论, 只打算在下一段叙述一下 Cramer 的结果. 这里我们提供一个反例, 说明极大似然估计确实可以不相合.

例 2.27 设总体分布为

$$\begin{cases} P_\theta(X=1)=1-P_\theta(X=0)=p, \text{ 当 } 0\leqslant p\leqslant 1, p \text{ 为有理数;} \\ P_\theta(X=1)=1-P_\theta(X=0)=1-p, \text{ 当 } 0\leqslant p\leqslant 1, p \text{ 为无理数.} \end{cases} \tag{2.51}$$

$X_1, \cdots, X_n$ 为抽自此总体的简单随机样本. 似然函数为

$$\begin{cases} p^T(1-p)^{n-T}, \ p \text{ 有理;} \\ p^{n-T}(1-p)^T, \ p \text{ 无理.} \end{cases} \quad \left(T=\sum_{i=1}^n X_i\right) \tag{2.52}$$

注意到 T/n 为有理数, 易见当 $p=\hat{p}_n=T/n$ 时, 似然函数(2.52)达到最大值. 故 $\hat{p}_n=T/n$ 即为 p 的极大似然估计. 但由总体分布的形式(2.51)知, $\hat{p}_n$ 以概率 1 收敛于 p 或 $1-p$, 视 p 为有理数或无理数而定. 因此, $\hat{p}_n$ 并非对每个 $p\in[0, 1]$ 都以概率 1 收敛于 p, 故不是 p 的相合估计.

(二) 相合渐近正态估计(CAM 估计)

定义 2.8 沿用定义 2.7 的记号. 若存在与样本大小 n 有关的、定义于 Θ 上的函数 $A_n(\theta)$ 和 $B_n(\theta)$, 其中 $B_n(\theta)$ 在 Θ 上处处大

于 0，使当 $n\to\infty$ 时

$$(\hat{g}(X^{(n)})-A_n(\theta))/B_n(\theta)\xrightarrow{\mathscr{L}}N(0,\ 1),\tag{2.53}$$

且 $\hat{g}(X^{(n)})$ 为 $g(\theta)$ 的弱相合估计，则称 $\hat{g}(X^{(n)})$ 为 $g(\theta)$ 的**相合渐近正态估计**，简称 **CAN 估计**(CAN 是 Consistent Asymptotic Normal 的缩写)．

就是说，CAN 估计是既相合，其分布又渐近于正态分布的那种估计．

本段我们要提出两个重要结果，即在很一般的条件下，矩估计为 CAN 估计，而在较有限制性的条件下，似然方程有一根为 CAN 估计．这些定理的证明比较复杂，要用到较多的极限理论知识，对于非数学专业的读者，在初读时可暂时放过这些证明．

1．矩估计　设样本 $X_1,\ \cdots,\ X_n$ 独立同分布，待估量 $g(\theta)$ 可表为(2.5)的形状．因为中心矩 μ_i 可表为前 i 个原点矩 $\alpha_1,\ \cdots,\ \alpha_i$ 的多项式，故可以将 $g(\theta)$ 表为只依赖于总体原点矩的形式，即

$$g(\theta)=G(\alpha_1,\ \cdots,\ \alpha_k).\tag{2.54}$$

取矩估计

$$\hat{g}=\hat{g}(X_1,\ \cdots,\ X_n)=G(a_{n1},\ \cdots,\ a_{nk}),\tag{2.55}$$

a_{ni} 的定义如(2.1)，再设总体的 $2k$ 阶原点矩存在且 G 对各变元有一阶连续偏导数．令

$$b_{ij}=\alpha_{i+j}-\alpha_i\alpha_j,\ i,\ j=1,\ \cdots,\ k;\ B=(b_{ij}).\ (k\text{ 阶方阵})\tag{2.56}$$

$$d_i=\partial G(\alpha_1,\ \cdots,\ \alpha_k)/\partial\alpha_i,\ i=1,\ \cdots,\ k;\ d=(d_1,\ \cdots,\ d_k)'.\tag{2.57}$$

$$b^2=d'Bd.\tag{2.58}$$

定理 2.6　在上述条件和记号下，当 $n\to\infty$ 时有

$$\sqrt{n}\,(\hat{g}(X_1,\ \cdots,\ X_n)-G(\alpha_1,\ \cdots,\ \alpha_k))\xrightarrow{\mathscr{L}}N(0,\ b^2).\tag{2.59}$$

证　我们先把证明所需的预备事实开列如下．这些事实都可以在一般的概率论教本中找到，或者可以由熟知的概率论结果容易地推出来：

a．(多维中心极限定理) 设 $Y_1, Y_2, \cdots$为一串独立同分布的 k 维随机向量，$EY_1=0$ 而 $\mathrm{Cov}(Y_1)=\Lambda$，则当 $n\to\infty$ 时，$\frac{1}{\sqrt{n}}(Y_1+\cdots+Y_n)$依分布收敛于 k 维正态分布 $N(0, \Lambda)$.

b．若当 $n\to\infty$ 时，$(Y_{n1}, \cdots, Y_{nk})$依分布收敛于 k 维正态分布 $N(0, \Lambda)$，而 $c_1, \cdots, c_k$ 都是常数，则当 $n\to\infty$ 时，$\sum_{i=1}^{k} c_i Y_{ni}$ 依分布收敛于一维正态 $N(0, c'\Lambda c)$，其中 $c'=(c_1, \cdots, c_k)$.

c．若随机变量 Y_n 依分布收敛于分布 F，而 Z_n 依概率收敛于 0，则 Y_n+Z_n 依分布收敛于 F.

d．若当 $n\to\infty$ 时，k 维随机向量$(Y_{n1}, \cdots, Y_{nk})$依分布收敛(于某一分布)，而 $Z_{n1}, Z_{n2}, \cdots, Z_{nk}$ 都依概率收敛于 0，则当 $n\to\infty$ 时，$\sum_{i=1}^{k} Z_{ni} Y_{ni}$ 依概率收敛于 0.

转到定理的证明．有

$$\hat{g}-G(\alpha_1, \cdots, \alpha_k)=\sum_{i=1}^{k} d_i(a_{ni}-\alpha_i)+\sum_{i=1}^{k} \varepsilon_{in}(a_{ni}-\alpha_i),$$

此处 $\varepsilon_{in}=\partial G(t_1, \cdots, t_k)/\partial t_i\big|_{(t_1,\cdots,t_k)=\varphi_n}$，$i=1, \cdots, k$，而 φ_n 是以$(\alpha_1, \cdots, \alpha_k)$和$(a_{n1}, \cdots, a_{nk})$为端点的线段内的某一点．因为 a_{ni} 为 α_i 的相合估计，且 G 的偏导数在$(\alpha_1, \cdots, \alpha_k)$点连续，有

$$\varepsilon_{in}\xrightarrow{P}0 \quad 当\ n\to\infty,\ i=1, \cdots, k. \tag{2.60}$$

若记 $Y_i=(X_i-\alpha_1, X_i^2-\alpha_2, \cdots, X_i^k-\alpha_k)$，$i=1, 2, \cdots$，则 $Y_1, Y_2, \cdots$独立同分布，且 $EY_1=0$，$\mathrm{Cov}(Y_1)=B$(B 见(2.56)式)．故依预备事实 a 有 $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\xrightarrow{\mathscr{L}}N(0, B)$．但易见 $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i=\sqrt{n}(a_{n1}-\alpha_1, \cdots, a_{nk}-\alpha_k)$，故得

$$\sqrt{n}(a_{n1}-\alpha_1, \cdots, a_{nk}-\alpha_k)\xrightarrow{\mathscr{L}}N(0, B). \tag{2.61}$$

这与预备事实 b 结合，即得

$$\sqrt{n}\sum_{i=1}^{k} d_i(a_{ni}-\alpha_i)\xrightarrow{\mathscr{L}}N(0, b^2). \tag{2.62}$$

b^2 见(2.58)．由(2.60)，(2.61)，及预备事实 d，知

$$\sqrt{n}\sum_{i=1}^{k}\varepsilon_{in}(a_{ni}-\alpha_i)\xrightarrow{P}0. \tag{2.63}$$

把(2.62)与(2.63)结合，并利用预备事实 C，即得(2.59). 定理证毕.

在不少情况下，要估计的 $g(\theta)$ 可表为一、两个中心矩的函数，比较简便，而要通过原点矩表出则很复杂. 因此，有必要给出在这种情况下的渐近正态性结果. 一般地，我们来考察形如

$$g(\theta)=H(\alpha_1,\ \mu_{t_2},\ \cdots,\ \mu_{t_c}), \tag{2.64}$$

其矩估计量为 $H(\overline{X}_n,\ m_{nt_2},\ \cdots,\ m_{nt_c})$. 使用与定理 2.6 的证明基本上相同的方法，但稍微复杂一些(复杂之处在于，中心矩不是独立同分布变量和)，可以证明下面的结果.

定理 2.6′ 设(2.64)式中的函数 H 在$(\alpha_1,\ \mu_{t_2},\ \cdots,\ \mu_{t_c})$点的邻域内有一阶偏导数，且此等偏导数在点$(\alpha_1,\ \mu_{t_2},\ \cdots,\ \mu_{t_c})$处连续. 则有

$$\sqrt{n}\{H(\overline{X}_n,\ m_{nt_2},\ \cdots,\ m_{nt_c})-H(\alpha_1,\ \mu_{t_2},\ \cdots,\ \mu_{t_c})\}$$
$$\xrightarrow{\mathscr{L}}N(0,\ b^2). \tag{2.65}$$

此处

$$b^2=\sum_{i=1}^{c}\sum_{j=1}^{c}\sigma_{ij}H_iH_j. \tag{2.66}$$

其中

$$H_1=\frac{\partial H}{\partial\alpha_1},\ H_i=\frac{\partial H}{\partial\mu_{t_i}},\ i=2,\ \cdots,\ c;$$

$$\sigma_{11}=\mu_2,\ \sigma_{1i}=\sigma_{i1}=\mu_{t_i+1}-t_i\mu_{t_i-1}\mu_2,\ i=2,\ \cdots,\ c;$$

$$\sigma_{ij}=\mu_{t_i+t_j}-t_i\mu_{t_i-1}\mu_{t_j+1}-t_j\mu_{t_i+1}\mu_{t_j-1}$$
$$-\mu_{t_i}\mu_{t_j}+t_it_j\mu_2\mu_{t_i-1}\mu_{t_j-1},\ i,\ j=2,\ \cdots,\ c. \tag{2.67}$$

如果 $g(\theta)$ 有 $H(\mu_{t_2},\ \cdots,\ \mu_{t_c})$ 的形状，即与 α_1 无关，则(2.65)式当然仍成立，只须把(2.66)式所确定的 b^2 修改为 $\sum_{i=2}^{c}\sum_{j=2}^{c}\sigma_{ij}\cdot H_iH_j$ 即可.

例 2.28 考察例 2.5. 被估计的量 $g(\theta)$ 为偏度 β_1、峰度 β_2 及变异系数 V. 其矩估计分别为

$$\hat{\beta}_1=m_{n3}/m_{n2}^{3/2},\ \hat{\beta}_2=m_{n4}/m_{n2}^2,\ \hat{V}=\sqrt{m_{n2}}/\overline{X}_n.$$

按(2.66)、(2.67)式，对这三个情况分别算得 b^2 值为：

$$b^2(\beta_1)=6,\ b^2(\beta_2)=24,\ b^2(V)=\frac{1}{2}V^2+V^4.$$

于是据定理 2.6′ 有

$$\sqrt{n}(\hat{\beta}_1-\beta_1)\xrightarrow{\mathscr{L}} N(0,\ 6),$$
$$\sqrt{n}(\hat{\beta}_2-\beta_2)\xrightarrow{\mathscr{L}} N(0,\ 24),$$
$$\sqrt{n}(\hat{V}-V)\xrightarrow{\mathscr{L}} N\left(0,\ \frac{1}{2}V^2+V^4\right). \tag{2.68}$$

值得注意的是，$\hat{\beta}_1$、$\hat{\beta}_2$ 的极限分布方差已与被估计的参数值无关. 这一点对 β_1、β_2 的大样本推断有用.

2. 极大似然估计　我们只考虑参数 θ 为一维的情况. 设总体有概率函数 $f_\theta(x)$. 为确定计，就设它是概率密度. 设参数空间 Θ 为一开区间 $(a,\ b)$. 假定 f_θ 满足以下各条件:

1° 设总体为 m 维. 在 R^m 中存在子集 $\mathscr{X}$，使对一切 $\theta\in\Theta$，有 $f_\theta(x)>0$ 当 $x\in\mathscr{X}$，$f_\theta(x)=0$ 当 $x\bar{\in}\mathscr{X}$.

2° 当 $x\in\mathscr{X}$ 时，$f_\theta(x)$ 对 θ 的前三阶导数在 Θ 上处处存在，且存在定义于 $\mathscr{X}$ 上的函数 $F_1(x)$，$F_2(x)$ 和 $H(x)$，使对一切 $\theta\in\Theta$ 和 $x\in\mathscr{X}$，有

$$|\partial f_\theta(x)/\partial\theta|\leqslant F_1(x),\ |\partial^2 f_\theta(x)/\partial\theta^2|\leqslant F_2(x),$$
$$|\partial^3\log f_\theta(x)/\partial\theta^3|\leqslant H(x), \tag{2.69}$$

其中

$$\int_{\mathscr{X}} F_i(x)dx<\infty,\ i=1,\ 2;\ \int_{\mathscr{X}} H(x)f_\theta(x)\,dx<\infty,\ \theta\in\Theta. \tag{2.70}$$

3° 对一切 $\theta\in\Theta$ 有

$$0<I(\theta)=\int_{\mathscr{X}}(\partial\log f_\theta(x)/\partial\theta)^2 f_\theta(x)dx<\infty. \tag{2.71}$$

定理 2.7　设 $X_1,\ \cdots,\ X_n$ 为取自满足上述条件 1°～3° 的总体的简单随机样本，且设对数似然方程

$$\sum_{i=1}^{n}\partial\log f_\theta(X_i)/\partial\theta=0 \tag{2.72}$$

有唯一根 $\hat{\theta}=\hat{\theta}(X_1,\ \cdots,\ X_n)$，则 $\hat{\theta}$ 是 θ 的 CAN 估计，且

$$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{\mathscr{L}} N\left(0,\ \frac{1}{I(\theta)}\right),\ \theta\in\Theta. \tag{2.73}$$

证 整个证明分两步. 第一步证明 $\hat{\theta}$ 相合, 第二步证明(2.73).

设 θ 的真值为 θ_0. 以 L 记 $\prod_{i=1}^{n} f_\theta(X_i)$. 把 $\partial\log f_\theta(X_i)/\partial\theta$ 作为 θ 的函数, 在 θ_0 点作 Taylor 展开, 有

$$\begin{aligned}\frac{\partial\log f_\theta(X_i)}{\partial\theta}=&\left.\frac{\partial\log f_\theta(X_i)}{\partial\theta}\right|_{\theta=\theta_0}\\&+(\theta-\theta_0)\left.\frac{\partial^2\log f_\theta(X_i)}{\partial\theta^2}\right|_{\theta=\theta_0}\\&+\frac{\varphi_i}{2}(\theta-\theta_0)^2H(X_i).\end{aligned}$$

此处 φ_i 为某一介于 0, 1 之间的量. 将上式对 $i=1, \cdots, n$ 相加, 可将似然方程(2.72)写为

$$0=\frac{1}{n}\frac{\partial\log L}{\partial\theta}=B_0+B_1(\theta-\theta_0)+\frac{1}{2}\varphi B_2(\theta-\theta_0)^2=h(\theta). \tag{2.74}$$

此处

$$B_0=\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial\log f_\theta(X_i)}{\partial\theta}\right|_{\theta=\theta_0},\quad B_1=\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2\log f_\theta(X_i)}{\partial\theta^2}\right|_{\theta=\theta_0},$$
$$B_2=\frac{1}{n}\sum_{i=1}^{n}H(X_i), \tag{2.75}$$

而 $0\leqslant\varphi\leqslant1$. 注意 φ 与 φ_i 一样, 与 θ_0、θ 和样本 $X_1, \cdots, X_n$ 都有关.

注意 B_0、B_1、B_2 都是 n 个独立同分布变量的算术平均. 现往证其均值都存在, 且

$$E_{\theta_0}B_0=0,\ E_{\theta_0}B_1=-I(\theta_0),\ E_{\theta_0}B_2=\int_{\mathscr{X}}H(x)f_{\theta_0}(x)dx=b_2. \tag{2.76}$$

最后一式由(2.70)立即得出, 且 $0\leqslant b_2<\infty$. 为证前两式, 注意由(2.69)和(2.70)知, $\int_{\mathscr{X}} f_\theta(x)dx$(其值为 1)可在积分号下对 θ 两次

求导，得

$$\int_{\mathscr{X}}(\partial f_\theta(x)/\partial\theta)\,|_{\theta=\theta_0}\,dx=\int_{\mathscr{X}}(\partial^2 f_\theta(x)/\partial\theta^2)\,|_{\theta=\theta_0}dx=0. \tag{2.77}$$

由(2.77) 得 $E_{\theta_0}B_0=\int_{\mathscr{X}}(\partial f_\theta(x)/\partial\theta)\,|_{\theta=\theta_0}\,dx=0$. 又

$$\begin{aligned}E_{\theta_0}B_1&=\int_{\mathscr{X}}(\partial^2\log f_\theta(x)/\partial\theta^2)\,|_{\theta=\theta_0}dx\\&=\int_{\mathscr{X}}(\partial^2 f_\theta(x)/\partial\theta^2)\,|_{\theta=\theta_0}dx\\&\quad-\int_{\mathscr{X}}\left(\frac{1}{f_{\theta_0}(x)}\frac{\partial f_\theta(x)}{\partial\theta}\bigg|_{\theta=\theta_0}\right)^2 f_{\theta_0}(x)dx\\&=-\int_{\mathscr{X}}(\partial\log f_\theta(x)/\partial\theta)^2|_{\theta=\theta_0}f_{\theta_0}(x)dx=-I(\theta_0),\end{aligned}$$

现任给 $\varepsilon>0$, 取 $\delta_1>0$ 充分小，其具体值待定. 由于 $0<I(\theta_0)<\infty$, $0\leqslant b_2<\infty$, 据大数定律，可找到只与 ε 和 δ_1 有关的自然数 n_0, 使当 $n\geqslant n_0$ 时，有

$$p_0=P_{\theta_0}(|B_0|\geqslant\delta_1)<\frac{\varepsilon}{3}, \tag{2.78}$$

$$p_1=P_{\theta_0}\left(B_1\geqslant-\frac{1}{2}I(\theta_0)\right)<\frac{\varepsilon}{3}, \tag{2.79}$$

$$p_2=P_{\theta_0}(B_2\geqslant b_2+1)<\frac{\varepsilon}{3}. \tag{2.80}$$

因此，若定义事件

$$B=\left\{|B_0|<\delta_1,\ B_1<-\frac{1}{2}I(\theta_0),\ 0\leqslant B_2<b_2+1\right\},$$

则当 $n\geqslant n_0$ 时，有 $P_{\theta_0}(B)>1-\varepsilon$.

现考虑(2.74)中的 $h(\theta)$. 当事件 B 发生时，有

$$h(\theta_0-\delta)>-\delta_1+\frac{1}{2}I(\theta_0)\delta-(b_2+1)\delta^2,$$

$$h(\theta_0+\delta)<\delta_1-\frac{1}{2}I(\theta_0)\delta+(b_2+1)\delta^2,$$

对任意 $\delta>0$. 现取 $\delta_1=\delta^2$, 并取 δ, 使

$$0<\delta<I(\theta_0)/(b_2+2),$$

则有$h(\theta_0-\delta)>0>h(\theta_0+\delta)$. 由条件1°、2°推知，$h(\theta)=\frac{1}{n}\frac{\partial \log L}{\partial\theta}$是$\theta$的连续函数，故在区间$(\theta_0-\delta, \theta_0+\delta)$内，有似然方程的根. 但据假定，似然方程只有唯一根，因此，这个根就是$\hat{\theta}$. 这样，我们证明了：当事件B发生时，有$|\hat{\theta}-\theta_0|\leqslant\delta$. 故当$n\geqslant n_0$时，有$P_{\theta_0}(|\hat{\theta}-\theta_0|>\delta)=1-P_{\theta_0}(|\hat{\theta}-\theta_0|\leqslant\delta)\leqslant 1-P_{\theta_0}(B)<\varepsilon$. 因为$\varepsilon>0$和$\delta>0$都可以给得任意小，证明了$\hat{\theta}\xrightarrow{P_{\theta_0}}\theta_0$，即$\hat{\theta}$为$\theta$的相合估计.

现转向证明(2.73). 用(2.74)式，注意$h(\hat{\theta})=0$，有

$$\sqrt{n}(\hat{\theta}-\theta_0)=\sqrt{n}B_0\Big/\left(-B_1-\frac{1}{2}\varphi B_2(\hat{\theta}-\theta_0)^2\right). \tag{2.81}$$

适已证明$\hat{\theta}-\theta_0\xrightarrow{P_{\theta_0}}0$，又$|\varphi|\leqslant 1$而$P_{\theta_0}(B_2\to b_2)=1$，故$-\frac{1}{2}\varphi B_2\cdot(\hat{\theta}-\theta_0)^2\xrightarrow{P_{\theta_0}}0$，因为$B_1\xrightarrow{P_{\theta_0}}-I(\theta_0)$，故知当$n\to\infty$时，有

$$(2.81)\text{右边的分母}\xrightarrow{P_{\theta_0}}I(\theta_0).$$

由此可知，当$\sqrt{n}(\hat{\theta}-\theta_0)$的极限分布，与

$$\sqrt{n}B_0/I(\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{I(\theta_0)}\left.\frac{\partial\log f_\theta(X_i)}{\partial\theta}\right|_{\theta=\theta_0} \tag{2.82}$$

的极限分布相同. 但(2.82)右边和号下的n项独立同分布，各有均值0及方差

$$\frac{1}{I^2(\theta_0)}\int_{\mathscr{X}}\left(\frac{\partial\log f_\theta(x)}{\partial\theta}\right)^2 f_\theta(x)\Big|_{\theta=\theta_0}dx=\frac{1}{I(\theta_0)}.$$

于是由独立同分布变量和的中心极限定理，知当$n\to\infty$时，(2.82)依分布收敛于$N\left(0, \frac{1}{I(\theta_0)}\right)$. 如上所述，这就证明了(2.73). 定理证毕.

(三) 最优渐近正态估计(RAM估计)和渐近效率

一个参数的CAN估计一般有很多. 拿正态总体$N(a, \sigma^2)$中

a 的估计来说，样本均值和样本中位数都是 CAN 估计. 因此，就存在一个比较优劣的问题.

设 $\hat{g}_i=\hat{g}_i(X_1, \cdots, X_n)$, $i=1, 2$, 是 $g(\theta)$ 的两个 CAN 估计，并假定具体地有

$$\sqrt{n}\,(\hat{g}_i-g(\theta))\xrightarrow{\mathscr{L}}N(0, \sigma_i^2(\theta)), \ \theta\in\Theta, \ i=1, 2, \tag{2.83}$$

而 $\sigma_1^2(\theta)<\sigma_2^2(\theta)$，则我们认为 $\hat{g}_1$ 优于 $\hat{g}_2$. 这可以作如下的解释：任给 $c>0$，问“$\hat{g}_i$ 与 $g(\theta)$ 的偏差不超过 $c/\sqrt{n}$”的概率有多大？依 (2.83)，近似地有

$$P_\theta(|\hat{g}_i-g(\theta)|\leqslant c/\sqrt{n})\approx\frac{1}{\sqrt{2\pi}}\int_{-c/\sigma_i(\theta)}^{c/\sigma_i(\theta)}e^{-x^2/2}\,dx.$$

由此可知，若 $\sigma_1(\theta)<\sigma_2(\theta)$，则当 n 充分大时，有 $P_\theta(|\hat{g}_1-g(\theta)|\leqslant c/\sqrt{n})>P_\theta(|\hat{g}_2-g(\theta)|\leqslant c/\sqrt{n})$，对任何 $c>0$，就是说，当样本大小 n 足够大时，使用 $\hat{g}_1$ 去估计 $g(\theta)$，较之使用 $\hat{g}_2$，更有可能得到较准确的估计. 正是在这个意义上，我们说 $\hat{g}_1$ 优于 $\hat{g}_2$.

(2.83)式中的 $\sigma_i^2(\theta)$ 称为估计量 $\hat{g}_i$ 的**渐近方差**. 渐近方差只可以理解为极限分布的方差，而不能理解为 $n\mathrm{Var}_\theta(\hat{g}_i)$ 的极限. 实际上，即使有(2.83)，$\hat{g}_i$ 可以根本没有方差(甚至也可以没有均值). 使用这个术语，可以说：两个 CAN 估计渐近方差小的好. 这与在样本大小 n 固定时，两个无偏估计，方差小的好相似.

于是自然地就产生找渐近方差最小的 CAN 估计的问题，这种估计称为**最优渐近正态估计**，简称 **BAN 估计**(BAN 是 Best Asymptotic Normal 的缩写). 这里可以把问题分两步提：

1. 找一切 CAN 估计的渐近方差的下确界.

2. 找出其渐近方差达到这个下确界的 CAN 估计，也就是 BAN 估计.

关于第一个问题，早在 Fisher 就猜测，这个下确界就是 Fisher 信息量的倒数 $1/I(\theta)$. Fisher 是从他对极大似然估计的研究作出这一猜测的(极大似然估计是最好的估计，其渐近方差应是最小的). 现在我们从 C-R 不等式的角度观察，也觉得这个猜测是可信的. 然而在五十年代，Hodges 举了一个反例，证明即使

在 $N(\theta, 1)$这样简单的情况，也可以作出 θ 的 CAN 估计，其渐近方差在指定的 θ 值处为 0. 这一来就打破了这个猜测，这个事实引起了一些学者的兴趣. 经过一些学者的研究，证明了 Fisher 的猜测，在一定限制性的意义下是对的(具体地说，要排除某些象 Hodges 例子中那种"病态性的"CAN 估计，只对剩下的 CAN 估计求渐近方差下界). 这些研究都过于专门，不宜在此叙述.

这样，在一定意义下可以把 BAN 估计定义为"其渐近方差等于 $1/I(\theta)$的 CAN 估计". 据此，再结合定理 2.7，就得到极大似然估计的一个重要性质：在一定条件下，若似然方程的解唯一，且这个解 $\hat{\theta}$ 是极大似然估计，则 $\hat{\theta}$ 是 BAN 估计.

也可以用另外的方式来表达这个结果. 为此，引进"渐近效率"的概念. 设 $\hat{g}$ 为 $g(\theta)$之一 CAN 估计，其渐近方差为 $\sigma^2(\theta)$，则称

$$ae_{\hat{g}}(\theta)=\frac{1}{I(\theta)}\Big/\sigma^2(\theta)=(I(\theta)\sigma^2(\theta))^{-1} \qquad (2.84)$$

为 $\hat{g}$ (在 θ 点)的渐近效率. 据此，极大似然估计有渐近效率 1. 至于矩估计，我们前已证明它们在很一般的条件下为 CAN 估计. 不过，除了几个常见的例子(在其中矩估计与极大似然估计重合)之外，矩估计的渐近效率一般都远低于 1. 通常人们说矩估计不如极大似然估计，大抵就是指这一点而言.

最后再举两个确定渐近效率的例子.

例 2.29 设 $X_1, \cdots, X_n \sim N(a, \sigma^2)$，要估计 a. 以 $\hat{g}$ 记样本中位数. 据(1.38)式，并注意到在此处有 $f(\xi_{1/2})=f(a)=\frac{1}{\sqrt{2\pi}\sigma}$，有

$$2\sqrt{n}\,\frac{1}{\sqrt{2\pi}\sigma}(\hat{g}-a)\xrightarrow{\mathscr{L}}N(0, 1),$$

或写为

$$\sqrt{n}(\hat{g}-a)\xrightarrow{\mathscr{L}}N\left(0, \frac{\pi}{2}\sigma^2\right),$$

渐近方差为 $\frac{\pi}{2}\sigma^2$. 在此处有 $I(\theta)=\frac{1}{\sigma^2}$，故按(2.84)式有

$$ae_g(\theta)=2/\pi. \tag{2.85}$$

(2.85)大体上可解释为：当 n 很大时，用 n 个样本的样本中位数去估计总体均值 a，相当于用 $2n/\pi$ 个样本的样本均值去估计 a.

例 2.30 设 $X_1, \cdots, X_n$ 为抽自 Cauchy 分布（密度函数为 $[\pi(1+(x-\theta)^2)]^{-1}$）的简单随机样本，要估计 a. 仍以 $\hat{g}$ 记样本中位数，与例 2.29 同样的方法证明，渐近效率为 $8/\pi^2$（习题 17）.

习　　题

1. 严格证明(2.19)是 $N(a, \sigma^2)$ 的参数 (a, σ^2) 的 MLE.

2. 若 $X=e^\xi$ 而 $\xi\sim N(a, \sigma^2)$，则X的分布称为对数正态分布．求出X的密度．设 $X_1, \cdots, X_n$ 是X的简单随机样本，求 a 和 σ^2 的矩和极大似然估计.

3. 若存在充分统计量，且样本 X 有概率函数，则 MLE 是充分统计量的函数.

4. 设总体密度为 $\frac{1}{2\sigma}\exp(-|x-a|/\sigma)$，$\sigma>0$ 和 a 为未知参数．设 $X_1, \cdots, X_n$ 为抽自此总体的简单随机样本，求 a 和 σ 的矩及极大似然估计.

5. 设 X_1, X_2, X_3 为抽自密度是 $\frac{1}{2}e^{-|x-\theta|}$ 的总体的简单随机样本，$X_{(1)}<X_{(2)}<X_{(3)}$ 为次序统计量，则 $X_{(2)}$ 是 θ 的 MLE．就本例证明：MLE 不一定是充分统计量（提示：用因子分解定理，证明不存在函数 $g(X_{(2)}, \theta)$ 和 $h(X_1, X_2, X_3)$，使 $\frac{1}{8}\exp\left(-\sum_1^3|X_i-\theta|\right)=g(X_{(2)}, \theta)h(X_1, X_2, X_3)$，也可以直接由充分统计量定义证：证明在给定 $X_{(2)}$ 时，$X_{(1)}$ 的条件分布与 θ 有关）.

6. 设 $X_1, \cdots, X_n$ 为从具有指数分布(1.8)的总体中抽出的简单随机样本．已知 $\bar{X}$ 为 $\frac{1}{\lambda}$ 的无偏估计．弄清：$1/\bar{X}$ 是否为 λ 的无偏估计.

7. 设 $X_1, \cdots, X_n\sim N(\theta, 1)$．证明：$g(\theta)=|\theta|$ 没有无偏估计（提示：利用 $g(\theta)$ 在 $\theta=0$ 处不可导）.

8. 设 $X_1, \cdots, X_n$ 是从具 Poisson 分布(2.7)中抽出的简单随机样本．(a)求 $e^{-2\theta}$ 的 UMVUE．说明这个估计不合理之处．给出一个看上去“合理”的估计．(b)要 $g(\theta)$（定义于 $0<\theta<\infty$）的无偏估计存在，充要条件是什么？求出 $g(\theta)$ 的 UMVUE（提示：若 $g(\theta)$ 的无偏估计存在，则必存在只依赖于 $T=\sum_1^n X_i$ 的无偏估计，写出 $g(\theta)=E_\theta\hat{g}(T)=\sum_{i=0}^\infty e^{-n\theta}\frac{(n\theta)^i}{i!}\hat{g}(i)$．由此证

明充要条件为：$g(\theta)$可展开为幂级数，收敛半径为∞．比较系数得 UMVUE)．

9．证明定理2.1之逆成立：若$\hat{g}$为$g(\theta)$的UMVUE，$\mathrm{Var}_\theta(\hat{g})<\infty$对任何$\theta\in\Theta$．则当$E_\theta\hat{l}=0$且$\mathrm{Var}_\theta(\hat{l})<\infty$对一切$\theta\in\Theta$成立时，必有$\mathrm{Cov}_\theta(\hat{g},\hat{l})=0$对一切$\theta\in\Theta$(提示：用反证法．若$\mathrm{Cov}_{\theta'}(\hat{g},\hat{l})\neq 0$对某个$\theta'\in\Theta$，则存在常数$\delta$，$|\delta|$充分小，使$\mathrm{Var}_{\theta'}(\hat{g}+\delta\hat{l})<\mathrm{Var}_\theta(\hat{g})$．因为$\hat{g}+\delta\hat{l}$为无偏估计，得到矛盾)．

10．求以下各情况下的UMVUE：(a)$X_1,\ \cdots,\ X_n\sim B(n,\ p)$，$g(p)=p^2$．(b)$X_1,\ \cdots,\ X_n\sim R(0,\ \theta)$，$g(\theta)=\mathrm{Var}_\theta(X_1)$．(c)$X_1,\ \cdots,\ X_n\sim N(a,\ \sigma^2)$，$\theta=(a,\ \sigma)$，$g(\theta)=a\sigma$．

11．设$X_1,\ \cdots,\ X_n\sim N(\theta,\ 1)$．求$\theta^2$的UMVUE．证明：此UMVUE达不到$C$-$R$不等式的下界．

12．设$X_1,\ \cdots,\ X_m\sim N(a,\ \sigma_1^2)$，$Y_1,\ \cdots,\ Y_n\sim N(a,\ \sigma_2^2)$，各样本$X_1,\ \cdots,\ X_m,\ Y_1,\ \cdots,\ Y_n$独立．(a)设$\sigma_2^2=\rho\sigma_1^2$而$\rho>0$已知．证明：$a$的UMVUE存在，求出它．(b)证明当$\rho$未知(即$\rho$也是参数)时，$a$的UMVUE不存在(提示：(a)中求得的UMVUE唯一且与ρ有关．同时，(a)中求得的UMVUE即使当ρ未知时也是a的无偏估计)．

13．设$X_1,\ \cdots,\ X_n\sim R(0,\ \theta)$，$\theta>0$，$\hat{\theta}_n$为$\theta$的MLE．证明：$\hat{\theta}_n$是$\theta$的强相合估计及任意阶的矩相合估计(提示：证强相合时要用到 Borel-Cantelli 引理：证明对任给$\varepsilon>0$有$\sum_1^\infty P(|\hat{\theta}_n-\theta|\geqslant\varepsilon)<\infty$)．

14．设$X_1,\ \cdots,\ X_n$是从一总体中抽出的简单随机样本，以θ记总体中位数．(a)利用(1.38)式证明：若总体有密度函数f，f在θ点大于0且连续，则样本中位数m_n是θ的弱相合估计．(b)利用强大数定律证明更强的结果：若总体分布只有唯一的中位数θ，则m_n是θ的强相合估计(提示：由θ唯一知，对任给$\varepsilon>0$有$F(\theta-\varepsilon)<1/2<F(\theta+\varepsilon)$，$F$为总体分布函数．由强大数律，以概率1成立{当$n$充分大时，$X_1,\ \cdots,\ X_n$中小于$\theta-\xi$者少于$n/2$个，大于$\theta+\xi$者也少于$n/2$个}，因此，以概率1成立：当$n$充分大时有$|m_n-\theta|\leqslant\varepsilon$)．

15．设$X_1,\ \cdots,\ X_n\sim N(a,\ \sigma^2)$，$S^2$为样本方差．证明$S^2$是$\sigma^2$的均方相合估计．

16．$\hat{\theta}_n$是θ的均方相合估计的充要条件为：$\hat{\theta}_n$渐近无偏，且$\lim\limits_{n\to\infty}\mathrm{Var}_\theta(\hat{\theta}_n)=0$，对任何$\theta\in\Theta$．

17．证明例2.30的结果．

18．设$X_1,\ \cdots,\ X_n\sim N(a,\ \sigma^2)$．在第一章(1.40)式中曾定义过$\sigma$的估计量$d$．证明(a)$d$非$\sigma$的无偏估计，但为渐近无偏．(b)修改$d$以得到$\sigma$的无偏估计$d^*$．(c)计算$d^*$的方差，因而证明$d^*$为均方相合．(d)证明$d^*$是

σ的强相合估计(提示：验证 $X_i-\overline{X}\sim N\left(0,\ \left(1-\frac{1}{n}\right)\sigma^2\right)$，由此不难算出 $E(d)$．为算 $\mathrm{Var}(d^*)$，要计算 $E|(X_i-\overline{X})(X_j-\overline{X})|$，$i=j$ 时容易．$i\neq j$ 时，先验证$(X_i-\overline{X},\ X_j-\overline{X})$为二维正态 $N\left(0,\ 0,\ \left(1-\frac{1}{n}\right)\sigma^2,\left(1-\frac{1}{n}\right)\sigma^2,\ -\frac{1}{n-1}\right)$．为证(d)，要利用强大数律，并利用估计式

$$\left|\sum_1^n|X_i-\overline{X}|-\sum_1^n|X_i-a|\right|\leqslant n|\overline{X}-a|).$$

第三章 假设检验

§3.1 概述 Pearson 和 Fisher 的思想

(一)引言

考察下面的简单例子.

例 3.1 有一大批产品要由工厂卖给商店，该批产品的废品率 p 未知，工厂与商店协商了一个数字 p'，例如 $p'=0.02$，约定当 $p\leqslant p'$ 时，商店接受这批产品，否则就拒收．由于 p 未知(且假定不可能进行全面检验以决定 p)，无法确切地知道是否该接受．于是双方同意用抽样检验的方法：从该批产品中随机地抽出 n 个，以 X 记其中废品个数，根据 X 的值，用某种双方同意的规则，去决定商店是否接受该批产品.

设 n 与整批产品个数相比很小，于是可足够准确地假定 X 服从二项分布 $B(n, p)$．这样，可以把上述纯由实用产生的问题归结为一个理论问题：样本 $X\sim B(n, p)$．要根据 X 的观察值，对命题 $H:p\leqslant p'$ 作出"是"或"否"的判断．在统计术语中，把这种需要根据样本去推断其正确与否的命题，称为**一个假设**或**统计假设**．通过样本以对一个假设作出"是"或"否"的判断的程序，称为**检验**这个假设，具体的判断规则称为该假设的**一个检验**．检验的结果若是肯定该命题，则称**接受**这个假设．其反面则是**否定**或**拒绝**这个假设．注意此处的"接受"和"否定"这种用语．它反映当事者在所面对的样本证据之下，对该命题所采取的一种态度、倾向性，以至某种必须或自愿采取的行动，而不是在逻辑上"证明"了该命题正确或不正确．这自然是因为样本有随机性．比方说．从一批废品率很小的产品抽出的样本中，也可能碰巧包含了较多的废品，而导致该批产品被拒收.

实用问题化为统计假设检验问题去处理，一般都与上述例子类似，可以概括为以下几条：

1．明确所要处理的问题．问题的回答只能是“是”或“否”．如例3.1，“是”就是“这批产品应予接受”．

2．设计适当的观察或试验以取得样本 X，X 的概率分布（常是通过这分布的参数或数字特征，如上例的 p）必须与问题有一种确定的联系．确切地说：知道了 X 的分布（或其参数、数字特征），就能明确无误地回答所提问题．

3．把问题的一种回答，例如“是”，作为一个命题．这个命题转化到样本 X 的分布上，就得到关于后者（或其参数、数字特征）的一个等价命题，称为一个假设．

4．根据样本 X 的具体值，按照一定的规则，作出接受或否定假设的决定．回到原问题，就等于回答为“是”或“否”．

假设检验是一种有重要应用价值的统计推断形式．从理论上说，它是数理统计学的一个重要分支．除去某些早期的片段情况不计，可以明确地说，假设检验方法和理论的系统发展，始于本世纪初．大致上按照时间顺序，其过程经历了以下一些重大事件：

(1) K. Pearson 的拟合优度 χ^2 检验（1900年）；

(2) R. A. Fisher 的显著性检验（本世纪二十年代）；

(3) J. Neyman 和 E. S. Pearson的理论（1928年开始）；

(4) A. Wald 的统计判决理论（1950年）；

(5) Bayes 方法．其中，Bayes 方法的确切年代不好定：这方法的基本观点可追溯到 T. Bayes(1702～1761)．本世纪早期已有不少学者鼓吹，但成为一种有影响的方法还要算在战后．上述4、5两项将在第五章中讨论，本章只涉及前三个题目．

Neyman–Pearson 关于假设检验的理论（NP 理论），是建立在概率的频率解释基础上的、关于假设检验的一套形式比较完美的数学理论，通过适当的表述方式，可以把 Pearson 和 Fisher 的工作包括到这个理论中去讨论．这样做在数学上较易处理，并为一些统计著作所采用．但作为一个初次接触假设检验这个题目的

读者来说，很有必要了解一下Pearson 和 Fisher 这些大师的思想，因为他们是从很实际的角度出发来考虑问题．通过对他们的工作的学习，对假设检验中一些基本概念的理解很有帮助，而这些正是初学者容易忽视的地方．从历史的角度说，NP 理论也不能算是另起炉灶，而是对他们的工作的继承和发展．以此之故，在这引论性的一节中，对 Pearson 和 Fisher 的观点作些介绍是有益的．

（二）Pearson 的思想．拟合优度检验

1938 年，K. Pearson 的儿子、本身也是著名统计学家的 E. S. Pearson，曾在一本关于他父亲的生平和工作的著作中，提到 K. Pearson 对统计的任务的看法是“To predict from past what will happen in the future”（从以往去预测将来会发生什么），以及在十九至二十世纪之交统计当务之急是“What was needed was a method for translating observed data into a predicative model”（需要的是一种方法，以将观察数据转化为一个可用于预测的模型）．Pearson 所谓“过去”，就是指已有的观察数据，“将来”则是指未来观察的可能结果．要做到由过去预测未来，必须用一个统计模型，确切地说就是一条分布（密度）曲线，去拟合已有的数据．然后用拟合的分布去计算在未来的观察中，出现种种值的可能性大小．为此 Pearson 提出了后来以他的名字命名的曲线系，希望在这个系统中，找出一条曲线，与已有的观测数据去进行拟合．在此需要处理以下两个问题：1．从曲线系中怎样去确定一条．2．估量拟合的程度如何．为解决第一个问题引出了他的矩估计法．关于第二个问题．Pearson 引进了一个统计量——χ^2 统计量（详见§3.2）$k=k(X_1, \cdots, X_n; F)$，以反映数据 $X_1, \cdots, X_n$ 及所拟合的分布曲线 F 之间的偏离．k 愈小拟合愈好．如果对一组具体观察数据算出 k 之值为 k_0，则根据 Pearson 在 1900 年证明的一个极限定理（定理 3.1），可以近似算出概率 $P(k\geqslant k_0)$．这个数可称为拟合优度．因为，这个数愈大（小），则产生象 k_0 这么大或更大的偏离的机会愈多（少），因而实际得到 k_0 这么大偏离这

件事并不稀奇(比较稀奇).

到此为止还没有假设检验的含义，只不过是定义了一个反映拟合程度优劣的指标——拟合优度. 如果必须作出“是否录用曲线 F”的决定,则必须定下一个阈值 α(如 $\alpha=0.01, 0.05, 0.1$ 等),规定当拟合优度小于 α 时,就不录用 F,否则就录用 F. 据此就可以把问题提成假设检验的形式:

$$\text{假设 } H: X_1, \cdots, X_n \text{ 是从具有分布 } F \text{ 的总体中抽出的样本.} \tag{3.1}$$

检验的方法是:指定阈值 α, 算出拟合优度 p. 当 $p<\alpha$ 时否定 H, $p\geqslant\alpha$ 时接受 H.

一般地，如果我们有一个理论、假说等，需要通过实践去检验之，则只要能设计一种观察或试验，使其结果 X 当理论或假说成立时有确定的分布 F. 则关于该理论或假说是否正确的问题，转化为检验假设(3.1)，其中 $X_1, \cdots, X_n$ 是 X 的独立观察值. 检验方法可以用上述 Pearson χ^2 检验法，也可以用种种不同的方法定义 $X_1, \cdots, X_n$ 与 F 之间的偏离,从而引出种种不同的检验法. 这些都可以称为拟合优度检验. 由 K. Pearson 开创的这个方向上的工作,是假设检验的一个重要组成部分.

(三) Fisher 的思想·显著性检验

1919 年,Fisher 进入 Rothamsted 农业试验站，那里有 K. Pearson 领导的一个统计学家小组,Fisher 在此从事统计学和遗传学方面的研究工作. 他通过田间试验研究试验设计,与此结合,对由试验数据作出归纳式推断的基本原理，作了探讨. 他的成果后来大都总结在他的名著«The Design of Experiments»中,此书前三章着重讨论了一些基本观点,其中包括他提出的显著性检验. 现通过 Fisher 仔细论述过的例子,来说明他的观点.

例 3.2 女士品茶的试验.

一种饮料由牛奶与茶按一定比例混合而成，可以先倒茶后牛奶(TM)或反过来(MT). 某女士声称，她可以鉴别是 TM 还是

MT. 设计如下的试验,来检验她的说法是否可信. 准备8杯饮料TM和MT各半,把它们随机地排成一列让该女士依次品尝,并告诉她TM和MT各有4杯. 然后请她指出那4杯是TM. 设她全说对了.

Fisher 推理过程如下:引进一个假设

$$H\text{: 该女士并无鉴别力.} \tag{3.2}$$

其意义是这样的:当 H 正确时,不论该女士如何做,她事实上只能从所提供的8杯中随机地挑选4杯作为TM. 从8杯中挑4杯,不同的挑法有 $\binom{8}{4}=70$ 种,其中只有一种是全部挑对. 因此,若该女士果真全部挑对,则我们必须承认,下述两个情况必发生其一:

1. H 不成立,即该女士确有一定的鉴别力;

2. 发生了一件,其概率只有 $\frac{1}{70}$ 的事件.

第二种情况相当于在一个盛有70个球的合子里随意摸出一个,正好摸到了事先指定的那一个. 这种情况比较希奇,因而有相当的理由承认第一种可能性. 或者说:该女士4杯全挑对这个结果,是一个不利于假设 H 的显著的证据. 据此,我们否定 H. 这样一种推理过程就叫做显著性检验.

如果该女士只说对了3杯,则表面上看,4杯说对3杯,成绩不错. 但我们要计算一下,纯粹出于碰巧而得到这个以至更好的成绩,其机会有多大. 简单计算表明:在 H 成立,即70种不同挑法为等可能时,挑中杯数 $\geqslant 3$ 的概率为 $17/70=0.243$. 发生一个概率为0.243的事件并不希奇,因此,试验结果没有提供不利于 H 的显著证据.

自然,人们可以说,$\frac{1}{70}$ 的概率虽然不大,但在一次试验中发生了总非不可能. 这个说法无法驳倒. 人们对这种问题的态度,与事情的重要性及可能的后果有关. 要得到一个判断的决定,就

须指定一个阀值α(0.01, 0.05, 0.1等). 只有在算出的概率$\left(\text{即上文的}\ \frac{1}{70}, 0.243\ \text{等}\right)$小于$\alpha$时, 才认为结果是显著的(提供了不利于$H$的显著证据),并导致否定$H$. 如在此例中,当取$\alpha=0.01$时,即使4杯全对也不认为结果显著, 而若取$\alpha=0.05$, 则认为是显著了. 读者不难理解:这里并无任何矛盾,因为α值是约定的. *α称为检验所用的显著性水平. α愈低, 获得显著结果愈难, 所导致的否定H的结论愈觉可信.*

据这个例子不难把Fisher显著性检验的思想归纳成以下几点:

1. 有一个明确的命题(假设)H.

2. 设计一定的试验,观察某变量X. X要有这样的性质:当H成立时,X有已知的分布. 如在上例中,若以X记该女士说对的杯数,则X有超几何分布: $P(X=i)=\binom{4}{i}\binom{4}{4-i}\Big/\binom{8}{4}$, $i=0, 1, \cdots, 4$.

3. 根据H和X的具体内容, 对X的值排一个次序. 使愈靠前的值,愈对H不利. 在上例中,这个排法是4, 3, 2, 1, 0.

4. 以x记X的观察值. 按2中求出的分布, 把x和比x更靠前的值的概率之和求出来, 暂记为p_x. p_x愈小, 试验结果x愈不利于H.

5. 算出p_x后,统计学家的工作也就完了. 他可以把p_x(连同其得出过程)报告给主事人,让后者去估量它的含义. 如果必须提出明确结论, 则必须事先给出(由主事者自定或有关主事者协商)显著性水平α,当$p_x<\alpha$时否定H, $p_x\geqslant\alpha$时接受H.

显著性检验这个名词还可以从更实际的意义上去理解. 在工农业中,假设检验问题常在这种情况下出现. 有两个处理(如两种工艺流程、两个种子品种、两种施肥方法等)A、B. A是原有的,已用了相当一个时期. B是新的,设想是一种改进. 我们要通过试验来判断B是否确优于A,于是引进假设H:

$$H: A,\ B \text{ 效果一样}. \tag{3.3}$$

我们心目中当然是想否定 H. 只有试验结果表明 B 的优势是"显著"时，才有根据这样做. 何谓显著?就必须按前述的观点去解释，下面的例子(也出自 Fisher)清楚说明了这一点.

例 3.3 为比较 A、B 两种施肥方法何者为优，选择 15 块一般大的地，把每块分成形状大小一样的两小块，随机地将其中一块派给 A，另一小块给 B. 各小块产量为

块号															
A	188	96	168	176	153	172	177	163	146	173	186	168	177	184	96
B	139	163	160	160	147	149	149	122	132	144	130	144	102	124	144
$A-B$	49	-67	8	16	6	23	28	41	14	29	56	24	75	60	-48

算出 $\sum(A-B)=314$. 现在要在假设$\{H: A,\ B \text{ 效果一样}\}$之下，把可能的试验结果按对 H 不利的程度排队. 在 H 成立时，每块内 $A-B$ 值(即 49, $-67\cdots$等)，并非由于 A、B 效果不同，而是由于两小块的差别. 但随机化的结果，每一小块有同等可能$\left(\frac{1}{2}\right)$分给 A 或 B. 因此，如在第一块，依随机化结果不同，$A-B$ 可以是 49，也可以是 -49，要看较好的那小块派给 A 还是 B. 这样一来，这个试验的全部可能的 $\sum(A-B)$ 值有 2^{15} 个:

$$\pm(49)\pm(-67)\pm(8)\cdots\pm(60)\pm(-48), \tag{3.4}$$

实际得出的 314 是 2^{15} 中的一种. 当 $A,\ B$ 效果有较大差别时，$|\sum(A-B)|$应取大值. 于是，按其绝对值大小，把(3.4)中的 2^{15} 个值排成一列:

$$x_1,\ x_2,\ \cdots,\ x_{2^{15}} \tag{3.5}$$

(3.5)中的 2^{15} 个值，在 H 成立的前提下，为等可能，即每个出现的概率都是 2^{-15}. 找出 m，使 $314=x_m$. 则 314 及比之对 H 更不利的值，在 H 成立时出现的机会只有 $p_{314}=m|2^{15}$. 本例可具体算出$p_{314}<0.0001$. 因此，即使在 $\alpha=0.0001$ 的显著性水平下，也有理由否定 H. 这样，试验结果提供了很显著的证据，不利于 H. 由于 $314>0$，结论是: 有很显著的证据表明 A 优于 B.

Fisher 的理论中缺少一件很重要的东西：同一个假设可以用很多不同的方法去检验，如何比较其优劣？考虑这个问题，就需要建立合理的比较准则，并在这种准则之下找最优者. Fisher 没有考虑这个问题. 因此他虽在开创假设检验这个方向起了重大的作用，却未能建立一套形式完美的数学理论. 拿例 3.2 来说，可以把试验作一点修改. 不告诉该女士 TM 和 MT 各 4 杯，而让她一一指明 8 杯中的每一杯是 TM 或 MT，按她说对的杯数去下判断. 在直观上，读者一定会觉得，这修改后的设计比原来的好. 但在怎样的意义下好，如何确切地论证，并不是一目了然的. J. Neyman 和 E. S. Pearson 的理论补足了这一点.

§3.2 拟合优度检验

(一)引言

拟合优度检验问题的提法，已在§3.1 中说明过了：设有一个可观察的、一维或多维的随机变量 X. $X_1, \cdots, X_n$ 是 X 的独立观察值，F 是一个已知的分布函数，其维数与 X 的维数相同. 要利用样本 $X_1, \cdots, X_n$ 去检验假设

$$H: X \text{ 的分布为 } F. \tag{3.6}$$

或者也可以这样提：如果用分布 F 去**拟合**样本 $X_1, \cdots, X_n$，则拟合的优良程度如何. F 常称为**理论分布**.

在§3.1(二)中概述了处理这个问题的一般原则：设法确定一个量 $D(X_1, \cdots, X_n; F)$，它有某种理由可以作为样本 $X_1, \cdots, X_n$ 与 F 的偏离的度量. 就具体样本算出 D 之值，记为 D_0. 然后在假设 H 成立的条件下算出概率

$$p(D_0) = P(D \geqslant D_0 | H).$$

它称为在选定的偏离指标 D 之下，样本与理论分布的**拟合优度**. 这个数介于 0, 1 之间. $p(D_0)$愈大，表示样本与理论分布的拟合愈好，而假设(3.6)愈可信. 一般，事先根据某种考虑，给定一个介于 0, 1 之间(通常很小)的阈值 α，然后：

当 $p(D_0)<\alpha$ 时否定 H. 当 $p(D_0)\geqslant\alpha$ 时接受 H. (3.7)

这种类型的检验通称为**拟合优度检验**. 由于 D 可以用种种不同的方法定义,可以有种种不同的拟合优度检验. 其中最著名的,是 K. Pearson 在 1900 年提出的 χ^2 检验, 和 Колмогоров 在 1933 年提出的一种检验. 本节将讨论这些检验,以前者为主.

(二) Pearson χ^2 检验: 理论分布完全已知的情况

1. 先设 X 只取有限个不同的值 $a_1, \cdots, a_r$. 理论分布 F 集中在 a_i 点的概率记为 p_i. 假设(3.6)写为:

$$H: P(X=a_i)=p_i,\ i=1,\ \cdots,\ r.\ (p_i>0\ \text{已知}, \sum_1^r p_i=1) \tag{3.8}$$

以 ν_i 记 $X_1, \cdots, X_n$ 中等于 a_i 的个数. ν_i 称为 a_i 的**观察频数**. 有 $\sum_1^r \nu_i=n$. np_i 称为 a_i 的**理论频数**,意即当 X 的分布确为 F 时, ν_i"在理论上"应取之值 (事实上这时有 $E\nu_i=np_i$). K. Pearson 引进如下的统计量(常称 Pearson χ^2 统计量), 以反映样本与理论分布的偏离:

$$k=k(X_1,\ \cdots,\ X_n;\ F)=\sum_{i=1}^r(\nu_i-np_i)^2/np_i. \tag{3.9}$$

其直观背景是: $E(\nu_i-np_i)^2$ 当 H 成立时等于 $np_i(1-p_i)$, 而当 H 不成立时, $E(\nu_i-np_i)^2$ 一般要大于 $np_i(1-p_i)$. np_i 是一个调整因子.

如前所述, 对一组具体样本算出 $k=k_0$ 后, 要计算拟合优度 $p(k_0)=P(k\geqslant k_0|H)$. k 的真确分布很复杂,故精确地计算 $p(k_0)$ 不易. K. Pearson 在其 1900 年工作中证明了下面的重要定理:

定理 3.1(K. Pearson) 若假设 H 真确, 则当样本大小 $n\to\infty$ 时,统计量 k 的分布收敛于 χ^2_{r-1}, 即自由度为 $r-1$ 的 χ^2 分布.

由此定理,当 n 较大时可得 $p(k_0)$ 的近似值

$$p(k_0)\approx\left[2^{\frac{r-1}{2}}\Gamma\left(\frac{r-1}{2}\right)\right]^{-1}\int_{k_0}^{\infty}e^{-\frac{x}{2}}x^{\frac{r-3}{2}}dx. \tag{3.10}$$

此值可在较仔细的χ^2分布表上查出(必要时插值). 一般,先给定了一个值α(0.01, 0.05 等), 从χ^2分布表上查出满足条件$P(\chi^2_{r-1}\geqslant\chi^2_{r-1}(\alpha))=\alpha$的值$\chi^2_{r-1}(\alpha)$, 然后视$k_0>\chi^2_{r-1}(\alpha)$或否, 决定否定或接受$H$. 这就是 Pearson 的拟合优度$\chi^2$检验.

我们把定理 3.1 的证明放在 § 3.2 的(六)段. 现举一例.

例 3.4 一家工厂分早中晚三班, 每班 8 小时. 近期发生了一些事故, 怀疑班次不同与事故出现率是否有关. 在记录的近期 15 次事故中, 有 6 次在早班, 3 次在中班, 6 次在晚班. 要根据这一观测资料来作判断.

提出这样一个假设H:"事故的可能性大小与班次无关". 先要把这一陈述化为一个概率命题. 虚设一个变量X, 取 1、2、3 为值. 意义是: 若事故在早班发生, $X=1$; 在中、晚班发生, 则X相应等于 2 和 3. 这样, 上述假设H可写为(3.8)的形式:

$$F(\{i\})=1/3,\ i=1,\ 2,\ 3.$$

本例中$n=15$, $p_1=p_2=p_3=1/3$, $\nu_1=6$, $\nu_2=3$, $\nu_3=6$. 按(3.9)算出$k_0=\frac{1}{5}[(6-5)^2+(3-5)^2+(6-5)^2]=1.2$. 查$\chi^2$分布表(自由度为$3-1=2$), 知在$H$成立时, $P(k\geqslant k_0)>\frac{1}{3}$. 这个数并不很小(在一次试验中, 发生一个概率为$\frac{1}{3}$的事件, 并不希奇). 因此, 所掌握的资料并未给"事故可能性大小与班次有关"的说法给予充分支持. 自然, 这不能解释为资料证明了"无关", 甚至也不能解释为充分支持了"无关"的说法. 其确切意义只能是如上所述: 资料未给"有关"的说法以充分支持.

对统计思想不大熟悉的人, 对此可能会觉得难以理解. 因为看起来, 6:3:6 的比例应当是充分证明了中班事故率低于平均值$\frac{1}{3}$. 对此我们再加一点解释. 其实, 这里涉及一个纯推断或采取行动的问题. 我们所作的推断是说: 即使三班的事故率一样, 但仅由随机原因(正如把 15 个球随机放到三个合子中)产生这种表面差别的机会, 也大到 1/3 以上. 你在(平均)100 家事故率与班次

无关的工厂中，会发现 30 家以上，其表面差别甚至比这更大．因此从纯学理(推断)的角度，在这样的情况下断言事故率与班次有关，确不甚令人信服．但是，如果你是这工厂的一位工程师，你可能会认为，上述表面上的差别已构成充分的理由去作一番进一步的研究，看是否能找出中班事故率较低的原因．这样做也可能有益或无益．是否这样去做，上述推断中提供的数字 1/3 是一个考虑的因素．另外，事故后果的严重性，作进一步调查的费用等，也是在作出决策过程中的因素．

2．现设理论分布 F 为一般的．若 X 为一维，则选择适当的常数 $a_1, \cdots, a_{r-1}$，满足 $-\infty<a_1<\cdots<a_{r-1}<\infty$，以把 $(-\infty, \infty)$ 分解为 r 个区间：

$$I_1=(-\infty, a_1),\ I_2=[a_1, a_2),\ \cdots,$$
$$I_j=[a_{j-1}, a_j),\ \cdots,\ I_r=[a_{r-1}, \infty). \qquad (3.11)$$

若 X 是 m 维的而 $m>1$，则要把 R^m 分解为 r 个彼此无公共点的区域 $I_1, \cdots, I_r$．记

$$p_j=P_F(X\in I_j)=F(a_j)-F(a_{j-1}),\ j=1, \cdots, r.$$
$$(F(a_0)=0,\ F(a_r)=1) \qquad (3.12)$$

这些都是已知数，且 $a_1, \cdots, a_{r-1}$ 选择之使 $p_1>0, \cdots, p_r>0$．以 ν_j 记 $X_1, \cdots, X_n$ 中落在 I_j 内的个数，而作出统计量 k((3.9)式)．则在 H 成立(即 X 确有分布 F)时，当 $n\to\infty$ 时定理 3.1 的结论仍真．因此，象在 X 离散的情况一样，可计算拟合优度并对 H 加以检验．

这个作法，实质上就是用一个如下定义的离散分布 F^* 来代替 F：在 $I_1, \cdots, I_r$ 内各取一点 $a_1', \cdots, a_r'$．令分布 F^* 在 a_j' 点的概率为 p_j，$j=1, \cdots, r$．所算出的 k 值(3.9)，反映了样本 $X_1, \cdots, X_n$ 与分布 F^* 的偏离．不过我们可以这样想：F^* 是 F 的一种"粗糙化"，形象地说，在 F^* 中舍弃了 F 的某些"细部"．故若 $X_1, \cdots, X_n$ 与较粗的 F^* 的拟合尚且不好，自没有理由认为，它反面与更精细的 F 拟合得好．因此，若样本 $X_1, \cdots, X_n$ 与 F^* 的拟合优度值很低，则我们有充分理由认为"X 有分布 F"的假设不

对. 反过来则不然: 若 $X_1, \cdots, X_n$ 与较粗的 F^* 拟合不错, 它不一定能很好地拟合更精细的 F. 这些都是在选择(3.7)中的 a_j, 以及对结果作解释时, 要注意的问题.

一般, 若把(3.7)式中的 r 选得足够大, 并使 $a_1, \cdots, a_{r-1}$ 有适当的配置, 我们总可以把 F^* 与 F 弄到很接近, 以至在实用上认为满意. 但除非样本大小 n 很大. 这样一来每组的理论频数 np_i 和观察频数 ν_i 都会很小, 而 k 的真确分布与 χ^2_{r-1} 会有较大的距离. 以此, r 能取得多大, 取决于 n. 有一种经验法则, 认为每组的理论频数不应小于5. 另一点要注意的是: (3.7)式中的 $a_1, a_2, \cdots$ 等必须不依赖于样本. 就是说, 不能根据样本 $X_1, \cdots, X_n$ 的位置去选择它们, 而必须事先定好. 只有这样定理 3.1 的结论才有效. 在实际工作中这一点常没有被严格遵守. 在一般情况下, 由此而引起的误差大体上可认为能忽略不计.

(三) 理论分布带参数的情况

在许多问题中, 要检验的假设是: 变量 X 的分布属于一个确定的分布族 $\{F(x, \theta_1, \cdots, \theta_t): (\theta_1, \cdots, \theta_t) \in \Theta\}$. 确切地说, 有了 X 的独立观察值 $X_1, \cdots, X_n$, 要由它们去检验假设

$$H: \text{存在}(\theta_{10}, \cdots, \theta_{t0}) \in \Theta, \text{使 } X \text{ 的分布为} \\ F(x, \theta_{10}, \cdots, \theta_{t0}). \tag{3.13}$$

如在统计应用中, 常假定样本 $X_1, \cdots, X_n$ 抽自正态总体. 当对此有怀疑时, 可通过样本去检验. 在此, 并不要求 X 的分布为某一特定的正态分布, 例如 $N(0, 1)$, 而只要求存在 (a, σ), 使总体分布为 $N(a, \sigma^2)$. 问题当然也可从"拟合"的角度去提. 例如, K. Pearson 所考虑的, 从 Pearson 分布族中找一个分布去拟合已有的观察数据.

Pearson 在 1900 年的工作中讨论了这个问题, 方法是理论分布完全已知时的直接推广. 取 $I_1, \cdots, I_r$ 如(3.11). 记

$$p_j(\theta_1, \cdots, \theta_t) = F(a_j, \theta_1, \cdots, \theta_t) - F(a_{j-1}, \theta_1, \cdots, \theta_t). \tag{3.14}$$

表达式

$$k(\theta_1, \cdots, \theta_t)=\sum_{j=1}^{r}[\nu_j-np_j(\theta_1, \cdots, \theta_t)]^2/np_j(\theta_1, \cdots, \theta_t) \tag{3.15}$$

是 $X_1, \cdots, X_n$ 与分布 $F(x, \theta_1, \cdots, \theta_t)$偏离的度量. 但 $\theta_1, \cdots, \theta_t$ 为未知参数, Pearson 使用样本 $X_1, \cdots, X_n$ 去估计它们. 若 $\hat{\theta}_j=\hat{\theta}_j(X_1, \cdots, X_n)$为 θ_j 的适当估计, $j=1, \cdots, t$, 则以之取代(3.15)中的 θ_j, $j=1, \cdots, t$, 得统计量

$$k^*=k(\hat{\theta}_1, \cdots, \hat{\theta}_t). \tag{3.16}$$

Pearson 在 1900 年工作中认为, 在这个带参数的情况下, 定理 3.1 的结论仍真. 即当(3.13)的假设 H 成立而 $n\to\infty$ 时, k^* 依分布收敛于 χ^2_{r-1}. 若果如此, 则前述计算拟合优度的近似计算以及假设 H 的检验法, 都可照搬过来. 在二十年代, Fisher 发现了 Pearson 在推理中的疏忽, 指出自由度不是 $r-1$ 而应为 $r-1-t$. 而且, 关于估计量 $\hat{\theta}_j$ 的大样本性质还有所要求(比方说, 矩估计一般不行). 这个问题经 Fisher、Neyman、E. S. Pearson 等一些著名学者研究过. 例如证明了: 在一定条件下, 下面两种估计方法都适合要求(简记 $p_j(\theta_1, \cdots, \theta_t)$为 p_j)

1. 最小 χ^2 法　找 $\hat{\theta}_j=\hat{\theta}_j(X_1, \cdots, X_n)$, $j=1, \cdots, t$, 使

$$k(\hat{\theta}_1, \cdots, \hat{\theta}_t)=\min\{k(\theta_1, \cdots, \theta_t):(\theta_1, \cdots, \theta_t)\in\Theta\}. \tag{3.17}$$

实际计算中是通过解方程组

$$\partial k(\theta_1, \cdots, \theta_t)/\partial\theta_i=0, \text{即} \sum_{j=1}^{r}\left(\frac{\nu_j-np_j}{p_j}+\frac{(\nu_j-np_j)^2}{2np_j^2}\right)\frac{\partial p_j}{\partial\theta_i}=0,$$

$$i=1, \cdots, t.$$

这个方法的直观意义很明显: $F(x, \hat{\theta}_1, \cdots, \hat{\theta}_t)$是分布族$\{F(x, \theta_1, \cdots, \theta_t):(\theta_1, \cdots, \theta_t)\in\Theta\}$中与样本偏离最小的分布(当偏离按(3.15)定义时). 但这个方程组很难解.

2. 另一个方法是把$(\hat{\theta}_1, \cdots, \hat{\theta}_t)$定义为方程组

$$\sum_{j=0}^{r}\frac{\nu_j}{p_j}\frac{\partial p_j}{\partial\theta_i}=0, \ i=1, \cdots, t \tag{3.18}$$

之解. 其意义可以从两方面去解释. 一是表达式 $p_1^{\nu_1}\cdots p_r^{\nu_r}$ 是观察结果$(\nu_1, \cdots, \nu_r)$的似然函数. 找 $\theta_1, \cdots, \theta_t$ 使这似然函数达到最大, 即得(3.18). 故$(\hat{\theta}_1, \cdots, \hat{\theta}_t)$是 $\theta_1, \cdots, \theta_t$ 在间接意义下(即先由样本 $X_1, \cdots, X_n$ 过渡到 $\nu_1, \cdots, \nu_r$, 再用之去估计 $\theta_1, \cdots, \theta_t$)的极大似然估计. 另一个解释是: (3.18)是由(3.17)中舍弃括号内第二项而得, 故也称修正最小 χ^2 法. 舍弃的根据是按大数定律 $\frac{\nu_j}{n}\approx p_j$, 故$(\nu_j-np_j)^2/n$ 相对于 ν_j-np_j 为无穷小. 方程组(3.18)比(3.17)简单些, 但往往也需要用数值方法求解.

现在给出极限定理的严格表述. 其证明超出本书范围之外, 故从略. 可参看陈希孺《数理统计引论》p.298, p.302～305.

定理 3.2 设下列条件满足:

1° Θ 为 R^t 中的开集, 存在$(\theta_{10}, \cdots, \theta_{t0})\in\Theta$, 使 X 的分布为$F(x, \theta_{10}, \cdots, \theta_{t0})$(即假设(3.13)成立).

2° 按(3.14)定义 p_j. 对 Θ 中任意两个不同的点$(\theta_{11}, \cdots, \theta_{t1})$和$(\theta_{12}, \cdots, \theta_{t2})$, 有

$$\sum_{j=1}^{r}|p_j(\theta_{11}, \cdots, \theta_{t1})-p_j(\theta_{12}, \cdots, \theta_{t2})|>0.$$

3° $\partial p_j(\theta_1, \cdots, \theta_t)/\partial\theta_i$ 在 Θ 内连续, $i=1, \cdots, t$, $j=1, \cdots, r$.

4° 令

$$I_{rs}(\theta_1, \cdots, \theta_t)=\sum_{i=1}^{r}\frac{1}{p_i(\theta_1, \cdots, \theta_t)}\frac{\partial p_i(\theta_1, \cdots, \theta_t)}{\partial\theta_r}\times\frac{\partial p_i(\theta_1, \cdots, \theta_t)}{\partial\theta_s},$$

$r, s=1, \cdots, t$, 而 $I(\theta_1, \cdots, \theta_t)$为一个 t 阶方阵, 其(r, s)元为 $I_{rs}(\theta)$, 则对任何$(\theta_1, \cdots, \theta_t)\in\Theta$, $I(\theta_1, \cdots, \theta_t)$的行列式不为 0.

5° $(\hat{\theta}_1, \cdots, \hat{\theta}_t)$为方程组(3.18)的解, 且是$(\theta_1, \cdots, \theta_t)$的弱相合估计.

则当样本大小 $n\to\infty$ 时, 由(3.15)、(3.16)式定义的统计量 k^* 依分布收敛于自由度 $r-1-t$ 的 χ^2 分布 χ^2_{r-1-t}.

由这个极限分布的形式看出：参数个数 t 必须小于 $r-1$.

例 3.5 在例 2.12 中，描述了基因 A,B,O 的频率 θ_1, θ_2, θ_3 ($\theta_3=1-\theta_1-\theta_2$)与血型 O、A、B、AB 的频率 p_1, …, p_4 的关系为(2.22)式. 这个关系是在一定的遗传学理论之下，且假设所考察的人群中，交配是随机的，才能成立. 为检验这些假设是否成立，可以用 χ^2 检验法. 设观察了该群中 n 个人的血型，其中 O、A、B 和 AB 分别有 n_1, …, n_4 人. 写出似然函数(2.23)，算出 θ_1, θ_2, θ_3 的极大似然估计(注意要在 $\theta_1+\theta_2+\theta_3=1$ 的约束下求)，记为 $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$. 算出

$$k^*=\frac{(n_1-n\hat{\theta}_3^2)^2}{n\hat{\theta}_3^2}+\frac{(n_2-n(2\hat{\theta}_1\hat{\theta}_3+\hat{\theta}_1^2))^2}{n(2\hat{\theta}_1\hat{\theta}_3+\hat{\theta}_1^2)}+\frac{(n_3-n(2\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2^2))^2}{n(2\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2^2)}+\frac{(n_4-2n\hat{\theta}_1\hat{\theta}_2)^2}{2n\hat{\theta}_1\hat{\theta}_2}.$$

此处 $r=4$, $t=2$. 按定理 3.2，在假设成立之下，k^* 有极限分布 χ_1^2.

在本例中，变量 X 是一个只取 4 个值(比方说 1, 2, 3, 4)的离散变量，4 个值分别相应于 4 种血型，故不须先按(3.11)的方式去离散化.

例 3.6 为要检验一组样本 X_1, …, X_n 是否从正态总体 $N(a,\sigma^2)$中抽出的，先按(3.11)分组. 有

$$p_j(a,\ \sigma)=\frac{1}{\sqrt{2\pi}\,\sigma}\int_{a_{j-1}}^{a_j}e^{-(x-a)^2/2\sigma^2}dx,\ j=1,\ \cdots,\ r$$

($a_0=-\infty$, $a_r=\infty$). 方程组(3.18)有形式

$$\sum_{j=1}^{r}\frac{\int_{a_{j-1}}^{a_j}(x-a)e^{-(x-a)^2/2\sigma^2}dx}{\int_{a_{j-1}}^{a_j}e^{-(x-a)^2/2\sigma^2}dx}=0,$$

$$\sum_{j=1}^{r}\frac{\int_{a_{j-1}}^{a_j}(x-a)^2e^{-(x-a)^2/2\sigma^2}dx-\sigma^2\int_{a_{j-1}}^{a_j}e^{-(x-a)^2/2\sigma^2}dx}{\int_{a_{j-1}}^{a_j}e^{-(x-a)^2/2\sigma^2}dx}=0.$$

k^* 的自由度为 $r-1-2=r-3$.

上述方程组并不容易解. 于是自然地会这样想：若用通常的

方法，如极大似然估计 $\bar{x}$ 以及 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 去估计 a 和 σ^2，则大为方便. 在别的问题中也存在这种情况，即方程组(3.18)很难解，但直接由样本 $X_1, \cdots, X_n$ 去求参数 $\theta_1, \cdots, \theta_t$ 的极大似然估计则较易. 然而，在理论上可以证明：若用 $\theta_1, \cdots, \theta_t$ 的极大似然估计 $\theta_1^*, \cdots, \theta_t^*$ 代替(3.15)中的$(\theta_1, \cdots, \theta_t)$而不用(3.18)的解，则所算出的 χ^2 统计量不一定有极限分布 χ^2_{r-1-t}. 更确切地，证明了在 n 很大时，真正的拟合优度值 $P(k(\theta_1^*, \cdots, \theta_t^*)\geqslant k_0|H)$介于$P(\chi^2_{r-1-t}\geqslant k_0)$与$P(\chi^2_{r-1}\geqslant k_0)$之间(后者较大). 按定理 3.2 算出之值为前者，而当无参数时，按定理 3.1 算出之值为后者. 因此，这个结果可以形象地解释为：用极大似然估计$\{\theta_j^*\}$而不用(3.18)的解$\{\hat{\theta}_j\}$，相当于把损失掉的自由度 t 挽回一部分. 问题在于，$k(\theta_1^*, \cdots, \theta_t^*)$的极限分布不再是 χ 分布.

(四)χ^2 方法用于检验独立性

每个人按其是否吸烟可分成两类，按其是否患癌症也可分为两类. 如要研究在某一群人(如一国的成年人，在某个行业工作的人)中，吸烟与患肺癌是否有关，则可从这一群人中随机抽取若干个，一一记录其是否吸烟和是否患癌症，用所得资料去进行统计分析.

这类问题在应用上碰得很多. 一般，有一个由大量个体构成的总体，每一个体上可量度两个属性指标：A, B. 指标 A 分 r 级：$A_1, \cdots, A_r$，而指标 B 分 s 级：$B_1, \cdots, B_s$. 从该总体中随机抽出 n 个个体，测得第 i 个个体的指标状况为(A_{r_i}, B_{s_i}), $i=1, \cdots, n$. 要依据这些资料，判断 A、B 两个指标是否有关. 关系的性质自然是多样化的，例如，一种情况是当个体的 A 指标处在较高级位时，其 B 指标也倾向于处在较高级位(可称为**正相关**).

形式地引进一个随机向量$X=(X^{(1)}, X^{(2)})$, $X^{(1)}$和$X^{(2)}$分别记同一个体的 A、B 指标的级，而第 i 个个体的观察结果记为$(X_i^{(1)}, X_i^{(2)})=X_i$，按上文的记号，$X_i=(r_i, s_i)$, $i=1, \cdots, n$. 如

果 n 相对于总体中的全部个体数很小，则可以把 $X_1, \cdots, X_n$ 看作是 X 的独立随机观察值. 而"指标 A, B 无关"则解释为"$X^{(1)}$与 $X^{(2)}$独立". 记

$$p_{ij}=P(X^{(1)}=i,\ X^{(2)}=j),\ i=1,\ \cdots,\ r,\ j=1,\ \cdots,s. \tag{3.19}$$

则由概率论可知，"$X^{(1)}$, $X^{(2)}$独立"的充要条件是：存在 $p_1^{(1)}, \cdots, p_r^{(1)}$ 和 $p_1^{(2)}, \cdots, p_s^{(2)}$, 都大于 0, $\sum_{i=1}^{r} p_i^{(1)}=\sum_{i=1}^{s} p_i^{(2)}=1$, 使得

$$p_{ij}=p_i^{(1)}p_j^{(2)},\ i=1,\ \cdots,\ r,\ j=1,\ \cdots,\ s. \tag{3.20}$$

把 $p_i^{(1)}$, $p_j^{(2)}$ 等视为参数，则假设

$$H: A,\ B \text{ 指标无关，即 } X^{(1)}、X^{(2)} \text{独立}. \tag{3.21}$$

有(3.13)那样的形式. 事实上，(3.19)~(3.20)定义了一个二维分布族，而(3.21)的假设 H 就是说，X 的分布在这个族内. 因此，可用(二)的方法来检验这个假设.

以 n_{ij} 记 $X_1, \cdots, X_n$ 中取(i, j)为值的个数，则观察结果可列成一张表如下. 这种表称为**列联表**，或 $\boldsymbol{r}\times\boldsymbol{s}$ **列联表**. 2×2 列联表又常称为**四格表**. 表中 $n_{i\cdot}=\sum_{u=1}^{s} n_{iu}$, $n_{\cdot j}=\sum_{v=1}^{r} n_{vj}$.

$x^{(2)}$ \ $x^{(1)}$	1	…	j	…	s	
1	n_{11}	…	n_{1j}	…	n_{1s}	$n_{1\cdot}$
⋮	⋮		⋮		⋮	⋮
i	n_{i1}	…	n_{ij}	…	n_{is}	$n_{i\cdot}$
⋮	⋮		⋮		⋮	⋮
r	n_{r1}	…	n_{rj}	…	n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	…	$n_{\cdot j}$	…	$n_{\cdot s}$	n

此处 X 为离散型的，其值域只有 rs 个点，每个点相当于一个 I_i. 写出$\{n_{ij}\}$的似然函数，为

$$L=\prod_{i=1}^{r}\prod_{j=1}^{s}(p_i^{(1)}p_j^{(2)})^{n_{ij}}.$$

其 $\log L$, 并在 $\sum p_i^{(1)}=\sum p_j^{(2)}=1$ 的约束下求极值. 简单计算表明：

极大值当 $p_i^{(1)}=\hat{p}_i^{(1)}$ 和 $p_j^{(2)}=\hat{p}_j^{(2)}$ 时达到，其中

$$\hat{p}_i^{(1)}=n_{i\cdot}/n,\ i=1,\ \cdots,\ r,\ \hat{p}_j^{(2)}=n_{\cdot j}/n,\ j=1,\ \cdots,\ s. \tag{3.22}$$

算出 χ^2 统计量之值

$$k^*=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-np_i^{(1)}p_j^{(2)})^2}{np_i^{(1)}p_j^{(2)}}\Bigg|_{p_i^{(1)}=\hat{p}_i^{(1)},\ p_j^{(2)}=\hat{p}_j^{(2)}}$$

$$=n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right). \tag{3.23}$$

自由度按定理 3.2 应为

$$(rs-1)-[(r-1)+(s-1)]=(r-1)(s-1). \tag{3.24}$$

因为 $p_1^{(1)},\ \cdots,\ p_r^{(1)}$ 中只有 $r-1$ 个独立参数，$p_1^{(2)}$、$\cdots$、$p_s^{(2)}$中只有 $s-1$ 个独立参数，全部独立参数个数为$(r-1)+(s-1)=r+s-2$. 指定阈值 α，查表找出 $\chi^2_{(r-1)(s-1)}(\alpha)$. 当(3.23)式算得的 k^* 超过 $\chi^2_{(r-1)(s-1)}(\alpha)$时，否定假设 H，即认为 A, B 两指标有关.

也可能，指标是连续取值而不是分成几个离散的级别. 例如，A 指标是一个人每日的平均运动时间，而 B 指标是其体重. 这时，可以把每个指标的取值范围象(3.11)那样分成若干个区间，然后按离散的情况去处理之. 所考虑的指标个数也可以多于 2. 这时称**多重列联表**. 独立性检验的方法是类似的.

顺便指出：当 $r=s=2$ 时，由(3.24)知自由度为 1. 按 Pearson 1900 年工作(认为自由度不因有参数而变化)则应为 3. Fisher 正是在这个特例中看出 Pearson 的错误. 他在 1922、1923 年的工作中，用实际数据给出了自由度应为 1 的、有说服力的证据.

(五) χ^2 方法用于检验齐一性

设有 r 个生产同一种产品的工厂，产品分为 s 个等级. 第 i 个工厂中的 j 等级品率为 $p_i(j)$, $j=1,\ \cdots,\ s$, $i=1,\ \cdots,\ r$. 两个工厂产品质量相同一语，理解为其相应等级品率相同. 于是"r 个工厂产品质量齐一"这个假设，可表为

$$H: p_1(j)=\cdots=p_r(j),\ j=1,\ \cdots,\ s. \tag{3.25}$$

现在从第 i 个工厂的产品中抽出 $n_{i\cdot}$ 个，记录其中的 j 等品有 n_{ij} 个，$j=1, \cdots, s$, $i=1, \cdots, r$. 要依据观测结果 $\{n_{ij}\}$ 去检验假设 H.

一般，有 r 个各包含大量个体的同类总体。每个个体的指标可处在 s 个等级中的一个，其余 $p_i(j)$, $n_{i\cdot}$, n_{ij} 等，都有相应的解释，则(3.25)的假设 H 就理解为"这 r 个总体(就这指标的分布而言)是齐一的"，因此 H 常称为齐一性假设.

所得数据 $\{n_{ij}\}$ 也可以象独立性检验那样排成一个 $r\times s$ 列联表. 不仅如此，形式上我们可以把齐一性假设看成是一种独立性假设. 为此对每一个体引进两个指标 $X^{(1)}$ 和 $X^{(2)}$. $X^{(1)}$ 是该个体所属总体的编号，而 $X^{(2)}$ 是它的指标等级，例如，对第二个工厂的三等品而言，$X^{(1)}=2$, $X^{(2)}=3$. 这时，"指标分布与总体编号无关"一语，就可以改说成"$X^{(1)}$ 与 $X^{(2)}$ 独立". 因此，可用(三)中的方法去检验假设(3.25)，所有的计算全相同.

但是，若仔细考察一下，就会看到二者之间有一个实质性的差别：在齐一性问题中，从第 i 总体中抽出的个体数 $n_{i\cdot}$ 是事先给定的，它没有随机性. 而在独立性问题中，相应的 n_i 是：样本中 A 指标处在等级 i 的个体数. 这个数事先并无规定而依赖于抽样的结果，是随机的. 由于有了这个差别，定理 3.2 就不能直接用于齐一性问题，而必须从头开始. 经过复杂的论证(见陈希孺《数理统计引论》p.310～315)，证明了在齐一性问题中，若假设(3.25)成立，则当 $n_{i\cdot}\to\infty$, $i=1,\cdots,r$ 时，由(3.23)定义的统计量 k^* 仍有极限分布 $\chi^2_{(r-1)(s-1)}$. 这才算是确实证明了，(三)中独立性检验的方法仍可用于此处.

如果所考察的指标是连续取值的，则以 F_i 记第 i 总体中指标的分布，而假设(3.25)写成

$$H: F_1\equiv F_2\equiv\cdots\equiv F_r. \qquad (3.26)$$

这种问题常称为多样本问题，有种种检验方法. χ^2 方法即其中之一. 为此，先象(3.11)那样，把指标的值域分组以实现离散化，并将 n_{ij} 定义为：第 i 总体所抽出的样本中，其指标值落在 I_j 内的个数. 然后用列联表的方式去处理.

(六)定理 3.1 的证明

沿用(3.8)、(3.9)式的记号. 注意在假设(3.8)成立时, $(\nu_1, \cdots, \nu_r)$有多项分布:

$$P(\nu_1=n_1, \cdots, \nu_r=n_r)=\frac{n!}{n_1!\cdots n_r!}p_1^{n_1}\cdots p_r^{n_r}$$

(当 $n_1, \cdots, n_r$ 都是非负整数, 且和为 n). 由此知$(\nu_1, \cdots, \nu_r)$的特征函数为(以下 $i^2=-1$)

$$g(t_1, \cdots, t_r)=(p_1e^{it_1}+\cdots+p_re^{it_r})^n.$$

令 $Y_j=(\nu_j-np_j)/\sqrt{np_j}$, $j=1, \cdots, r$. 则 $k=\sum_{j=1}^r Y_j^2$, 又$(Y_1, \cdots, Y_r)$有特征函数

$$\varphi(t_1, \cdots, t_r)=\exp\left(-i\sqrt{n}\sum_{j=1}^r\sqrt{p_j}\,t_j\right)\times g\left(\frac{t_1}{\sqrt{np_1}}, \cdots, \frac{t_r}{\sqrt{np_r}}\right),$$

取 log, 作 Taylor 展开, 易得

$$\begin{aligned}\log\varphi(t_1, \cdots, t_r)&=n\log\left[1+i\frac{1}{\sqrt{n}}\sum_{j=1}^r\sqrt{p_j}\,t_j-\frac{1}{2n}\sum_{j=1}^r t_j^2+O(n^{-3/2})\right]-i\sqrt{n}\sum_{j=1}^r\sqrt{p_j}\,t_j\\&=-\frac{1}{2}\sum_{j=1}^r t_j^2+\frac{1}{2}\left(\sum_{j=1}^r\sqrt{p_j}\,t_j\right)^2+O(n^{-1/2}).\end{aligned}$$

于是得

$$\lim_{n\to\infty}\varphi(t_1, \cdots, t_r)=\exp\left(-\frac{1}{2}Q(t_1,\cdots,t_r)\right), \qquad (3.27)$$

其中二次型 $Q(t_1, \cdots, t_r)=\sum_{j=1}^r t_j^2-\left(\sum_{j=1}^r\sqrt{p_j}\,t_j\right)^2$ 的方阵为 $A=I-pp'$, I 为 r 阶单位阵而 $p=(\sqrt{p_1}, \cdots, \sqrt{p_r})'$.

记 $Y=(Y_1, \cdots, Y_r)'$, 作正交变换

$$Z=(Z_1, \cdots, Z_r)'=BY,$$

使 B 的第一行为 p'. 则 $Z_1=p'Y=\frac{1}{\sqrt{n}}\sum_{j=1}^r(\nu_j-np_j)\equiv 0$. 故由

变换的正交性，有

$$k=Y_1^2+\cdots+Y_r^2=Z_1^2+\cdots+Z_r^2=\sum_{j=2}^{r} Z_j^2. \qquad (3.28)$$

Z 的特征函数为 $\psi(u)=\varphi(B'u)$, $u=(u_1, \cdots, u_r)'$. 由(3.27)，有

$$\lim_{n\to\infty} \psi(u)=\exp\left(-\frac{1}{2} Q(B'u)\right). \qquad (3.29)$$

但

$$Q(B'u)=(B'u)'I(B'u)-(B'u)'pp'(B'u)=u_2^2+\cdots+u_r^2. \qquad (3.30)$$

因此 $(B'u)'I(B'u)=u'BB'u=u'u$, 而 $p'B'$ 除第一元为1外，其他元为0(这由 B 正交且第一行为 p' 推出)，故 $p'(B'u)=u_1$, 因而 $(B'u)'pp'(B'u)=u_1^2$. 由(3.29), (3.30), 知

$$\lim_{n\to\infty} \psi(u)=\exp\left(-\frac{1}{2}(u_2^2+\cdots+u_r^2)\right). \qquad (3.31)$$

由此式及特征函数的连续性定理，知 $(Z_2, \cdots, Z_r)$ 当 $n\to\infty$ 趋于一个 $r-1$ 维分布，其各分量独立且都有分布 $N(0, 1)$. 因此当 $n\to\infty$ 时有 $\sum_{j=2}^{r} Z_j^2 \xrightarrow{\mathscr{L}} \chi_{r-1}^2$. 再注意到(3.28)，即得欲证的结果.

(七) Колмогоров 检验

这个检验的想法如下：先通过样本 $X_1, \cdots, X_n$ 对 X 的分布函数作一个估计. 若此估计接近于给定的分布函数 F, 就接受假设(3.6)，不然就否定(3.6). 这里设 X 为一维的.

定义 3.1 称定义在 $-\infty<x<\infty$ 的函数

$$F_n(x)=F_n(x; X_1, \cdots, X_n)=\frac{1}{n}(X_1, \cdots, X_n \text{ 中，小于 } x \text{ 的个数})$$

为 $X_1, \cdots, X_n$ 的**经验分布函数**.

注意，当固定样本 $X_1, \cdots, X_n$ 时，$F_n(x; X_1, \cdots, X_n)$ 作为 x 的函数，满足一维分布函数的三个基本性质：非降、左连续、$F_n(-\infty)=1-F_n(\infty)=0$. 事实上它是这样一个随机变量 ξ 的分布函数，ξ 取每个 X_i 为值的概率都是 $1/n$(若有 n 个 X_i 取同一值

a, 则 $P(\xi=a)=\frac{1}{n}(X_1, \cdots, X_n$ 中等于 a 的个数). 利用这一点不难把经验分布函数的概念平行地推广到 X 为多维的情况.

当固定 x 时, $F_n(x, X_1, \cdots, X_n)$ 作为 $X_1, \cdots, X_n$ 的函数, 是一个统计量, 具有性质:

$$E(F_n(x))=P(X<x), \quad \lim_{n\to\infty} F_n(x)=P(X<x), \ a.s..$$

由这些性质可知, F_n 是 X 的分布的一个良好估计. 定义一个反映经验分布 F_n 与假设的理论分布 F 的偏离的量. 这样的量当然有很多方法可以定义. Колмогоров 选择的是一致距离:

$$\|F_n-F\|=\sup\{|F_n(x)-F(x)|: -\infty<x<\infty\}.$$

它常称为 F_n 与 F 之间的Колмогоров 距离. 于是, 按在(一)中概述的一般原则, 在有了具体样本后, 先算出上述距离之值 D_0, 然后计算概率

$$\tilde{p}(D_0)=P(\|F_n-F\|\geqslant D_0|H).$$

它就是在 Колмогоров 距离之下, 样本 $X_1, \cdots, X_n$ 与理论分布 F 的拟合优度. 若指定一个阀值 α 而按(3.7)的作法去检验, 则须定出常数 $D_{n\alpha}$, 使

$$\tilde{p}(D_{n\alpha})=\alpha.$$

然后, 当 $D_0>D_{n\alpha}$ 时否定 H, 不然就接受 H. 这就是 Колмогоров 的拟合优度检验. 但经验分布函数的概念不是始自 Колмогоров.

为了要具体施行这个检验, 或算出拟合优度 $\tilde{p}(D_0)$, 需要知道 $\|F_n-F\|$ 在 H 成立的条件下的分布. 这个分布的形式极为复杂, 不便应用. 1933 年, Колмогоров 证明下面著名的极限定理: 若 $F(x)$ 在 $-\infty<x<\infty$ 处处连续, 则

$$\lim_{n\to\infty} P(\|F_n-F\|\geqslant\lambda/\sqrt{n})=\begin{cases}1-\sum\limits_{i=-\infty}^{\infty}(-1)^i e^{-2i^2\lambda^2}, & \lambda>0;\\ 1, & \lambda\leqslant 0.\end{cases}$$

用这个定理, 可以在 n 较大时, 近似地决定 $D_{n\alpha}$ 之值. 例如, 当 n 较大时, $D_{n,0.01}\approx 1.628/\sqrt{n}$, $D_{n,0.05}\approx 1.358/\sqrt{n}$. 当 n 较小时, 对若干 α 值, $D_{n\alpha}$ 可以从有关的统计表上查到.

作为(3.6)的两个不同的检验，Pearson χ^2 检验与 Колмогоров 的检验，其优劣比较如何？大体上可以这样说：在 X 为一维且理论分布完全已知时，Колмогоров 检验优于 χ^2 检验. 因为 1. χ^2 统计量之值依赖于把 $(-\infty, \infty)$ 分为 r 个区间的具体分法，包括 r 的选择和区间的位置，Колмогоров 距离 $\|F_n - F\|$ 则没有这个依赖性. 2. 一般说来，Колмогоров 检验的鉴别力略高，就是说，在 F 不是 X 的分布时，用 Колмогоров 检验较易发觉之.

另一方面，Pearson 的 χ^2 检验有其优点. 1. 当 X 是多维时，处理方法与一维完全一样，极限分布的形式也与这个维数无关. 2. 尤其重要的是：对于理论分布包含参数这个重要情况，理论上和方法上都与理论分布完全已知的情况相似，只是极限分布的自由度相应缩小一些而已. 在 Колмогоров 检验的情况则不然. 对每个具体的分布族(例如，检验理论分布属于正态分布族、指数分布族(1.8)等)，需要作特殊的处理，且难度很大，目前只对极个别分布族作了出来，其中尚不包括极端重要的正态分布族（对这个情况，用随机模拟的方法造了表）.

除了 Колмогоров 检验以外，经验分布函数在假设检验(不一定是拟合优度检验)中还有一些其他的应用，如 Cramer-von Mises 检验、Смирнов 检验等. 这些检验在实用上的意义较小，而其理论又涉及很复杂的极限定理. 这些定理的证明甚至在一般的非参数统计专著中都不给出，而必须查看原始文献. 因此这里都从略了.

§3.3 Neyman-Pearson 理论

从 1928 年开始的大约 10 年时间内，J. Neyman 和 E. S. Pearson 合作发表了一系列的论文，建立了假设检验的一种数学理论，通称为 **Neyman-Pearson 理论**(简称 NP 理论). 这是假设检验 B 至整个数理统计学中的一个重大事件. 本节的目的是介绍 NP 理论的基本概念，有些细节内容留待 §3.4 讨论.

(一)问题提法．原假设和对立假设

设有样本 X，取值于样本空间 $\mathscr{X}$．只知道 X 的分布属于一个分布族$\{F_\theta, \theta\in\Theta\}$(这一点作为模型假定，或可由具体问题的提法得到，总之它不是检验对象)．设 Θ_H 是 Θ 的一个非空真子集，则命题 $H: \theta\in\Theta_H$ 称为一个**假设**或**原假设**，也有称为**零假设、解消假设**的．命题 H 的确切含义是：存在一个 $\theta_0\in\Theta_H$ 使 X 的分布为 F_{θ_0}．记 $\Theta_K=\Theta-\Theta_H$，则命题 $K: \theta\in\Theta_K$ 称为 H 的**对立假设**，也有称为**备选假设**的．表述

$$H: \theta\in\Theta_H \leftrightarrow K: \theta\in\Theta_K \tag{3.32}$$

称为一个假设检验问题．其意义是：根据样本 X 的具体值，去判断 H 是否正确．或者说，在 H 和 K 中选择一个，分别称为接受 H 和否定 H(也称为拒绝 H)．

关于原假设的含义，与 Fisher 的提法并无不同．但这里明确提出了对立假设．这看来是一件一目了然的事，但以后将看到，正是由于明确了这个提法，就有可能补足 Fisher 理论中的一个空白，即可以对同一假设的不同检验的优劣进行比较．

这里的提法，是已经把实际问题数学化了．如果一开始问题并不是这样数学化了(象例 3.2, 3.3)，则必须先做这一步．这往往涉及对具体问题中一些提法如何解释，不一定是很容易的．观察以下的例子可知．

例 3.7 考虑例 3.2．以 X 记该女士说对的杯数．前已指出，在 H(该女士无鉴别力)成立之下，X 有超几何分布

$$P(X=i|H)=\binom{4}{i}\binom{4}{4-i}\Big/\binom{8}{4},\ i=0,\ 1,\ \cdots,\ 4. \tag{3.33}$$

但什么是对立假设？在对立假设下 X 的分布如何？这就取决于对"该女士有鉴别力"一语作何解释．一种合理的解释如下：以 p_i 记 $X=i$ 的概率．当该女士有鉴别力时，相对于(3.33)而言，对较大的 i，p_i 应较大些．这条件可表为

$$p_i+p_{i+1}+\cdots+p_4 \geqslant \sum_{j=1}^{4}\binom{4}{j}\binom{4}{4-j}\bigg/\binom{8}{4},$$

对 $i=1, 2, 3, 4$, 且不等号至少对一个 i 成立. (3.34)

这样就确定了对立假设下 X 的分布族. X 的分布族可视为依赖 4 个参数 $p_1, \cdots, p_4$(它们的和不超过 1). 当然, 也可以有别的解释. 比方说, 若该女士无鉴别力, 则应有 $EX=2$. 在她有鉴别力时, 平均说对杯数大于 2: $EX>2$. 把满足后一条件的分布算作对立假设中.

例 3.8 考虑例 3.3. 本例原假设和对立假设的确切提法都不容易.

把 15 块地编号: 1, 2, …, 15. 第 i 块分成两小块, 其条件有些差别. 可以把这个差别表示为: 在其他条件一样的情况下, 一小块单位面积产量比另一小块多 c_i. 又以 c 记由于处理 A、B 的不同而带来的单位面积产量差($c>0$ 表示 A 优于 B). 仍以 X 记 $\sum(A-B)$ 如例 3.3. 则 X 的分布为[1]: 以 2^{-15} 的概率取 $15c\pm c_1\pm c_2\pm\cdots\pm c_{15}$ 中的每一个值. 这个分布族包含 16 个参数 $c, c_1, \cdots, c_{15}$. 原假设为 $H: c=0, c_1, \cdots, c_{15}$ 任意, 对立假设为 $K: c\neq 0, c_1, \cdots, c_{15}$ 任意.

另一种定模型的方法如下: 假定每块内的两小块地力条件足够均匀. 但有些随机因素影响产量, 如耕作者在各小块上的操作不能完全一样, 气候因素对每小块的影响也不完全一致等. 仍以 c 记由于 A、B 不同所导致的单位面积产量差, 则第 i 小块上观察到的 $X_i=(A-B_i)$ 可表为 $c+e_i$ 的形式, e_i 是由上述随机因素所导致的误差. 假定 e_i 服从均值为 0 的正态分布, 则样本 $X_1, \cdots, X_n$ 独立, 各有分布 $N(c, \sigma^2)$, 原假设为 $H: c=0, \sigma>0$ 任意, 对立假设为 $K: c\neq 0, \sigma>0$ 任意. 由于出发点不同, 所导致的模型也就有根本差别. 在前一种提法中, 我们不要求两小块条件很一致, 但要求其他随机因素的影响可忽略不计, 数据的随机性纯由每块内两小块的随机分配而来. 在后一提法下, 由于假定了两小块条件很均

1) 为方便计, 就设每小块面积为一单位.

匀，它们的分配已不起什么作用．数据随机性来源于其他随机因素．

在确定原假设和对立假设时，要充分考虑和利用已知的背景知识．如把一物件在天平上秤 n 次得 $X_1, \cdots, X_n$，用以检验该物件重量是否为 1．设天平的随机误差服从正态分布 $N(0, \sigma^2)$．若对天平精度无所知，则检验问题为 $X_1, \cdots, X_n \sim N(a, \sigma^2)$，$H$：$a=1$（$\sigma$ 任意）$\leftrightarrow K$：$a \neq 1$（σ 任意）．若已知天平精度，则可认为 σ 已知，例如 $\sigma=0.1$，则检验问题为 $X_1, \cdots, X_n \sim N(a, 0.01)$，$H$：$a=1$，$K$：$a \neq 1$．又如在 § 3.2（三）的独立性 χ^2 检验问题中，若事先对指标 A、B 相关的可能情况无所知，则只能以一切不满足 (3.20) 的情况为对立假设．但往往事先对相关的可能情况有所知，例如，若不独立，则必是正相关．一个实例是吸烟与患肺癌的关系．二者或者无关．如有关，则必是吸烟使患肺癌的可能性增大．这时对立假设就有所限制，例如，反映正相关的一种提法是（仍用 § 3.2（三）的记号）

$$K\text{：当 } i<i' \text{ 时对 } j=1, \cdots, s \text{ 有 } \sum_{v=j}^{s} p_{iv} \leqslant \sum_{v=j}^{s} p_{i'v}.$$

且不等号至少对一组 $i<i'$ 及某个 j 成立．针对这种较确定的 K，可以提出比 § 3.2（三）中的 χ^2 法更为有效的检验方法．

（二）否定域，检验函数

设给了假设检验问题 (3.32)．这问题（或者说假设 H）的一个检验，就是一个这样的法则：一旦有了具体的样本，由该法则就可决定是应接受 H 还是否定 H．这样，一个检验等价于把样本空间 $\mathscr{X}$ 分解为两个不相交的部分 $\mathscr{X}_1$ 和 $\mathscr{X}_2$，使当样本属于 $\mathscr{X}_1$ 时接受 H，属于 $\mathscr{X}_2$ 时否定 H．$\mathscr{X}_2$ 称为该检验的**否定域**，也有称**临界域**的．$\mathscr{X}_1$ 则称为**接受域**．

例如，在例 3.1 中，一个合理的检验是有否定域 $\{X>c\}$，c 之值由买卖双方协商定之．例 3.2 中，可以取以 $\{4\}$ 为否定域的检验．例 3.3 的检验的否定域的仔细描述较复杂些，留给读者作为一个

习题(习题1).

在有些情况下，当样本取某些值 x 时，不是立即接受或否定 H，而是规定一个与 x 有关的概率 $\varphi(x)$. 再设计一个试验，使其中某事件 A 出现的概率为 $\varphi(x)$（例如，取一个 $[0, 1]$ 区间均匀随机数 ξ，定义 A 为"$0\leqslant\xi\leqslant\varphi(x)$"）. 具体作这个试验，若 A 发生了，则否定 H. 不然就接受 H. 如在例 3.1 中，买卖双方协商同意：若 $X>c$ 则买方拒收（否定 $p\leqslant p'$），若 $X<c$ 则接收. 当 $X=c$ 时，卖方觉得若规定拒收，则拒收可能性过大；若规定接受，则买方觉得，接收不合格产品批的可能性过大. 于是双方同意定下一个数 $\varphi(c)$，规定当 $X=c$ 时以概率 $\varphi(c)$ 拒收该批产品.

这样，可以把"检验"的概念推广为：检验是一个定义在样本空间 $\mathscr{X}$ 上的、取值于 $[0, 1]$ 区间的函数 $\varphi(x)$. $\varphi(x)$ 是当有了样本 x 时，否定 H 的概率，它称为**检验函数**. 有了样本 x 后，先算出 $\varphi(x)$. 若 $\varphi(x)=1$，则否定 H；若 $\varphi(x)=0$，则接受 H. 若 $0<\varphi(x)<1$，则还要安排一个随机试验，使其中某事件 A 的概率为 $\varphi(x)$. 视事件 A 是否发生，决定否定或接受 H. 若检验函数 φ 只取 0、1 为值，则它称为**非随机检验**，否定域就是 $\{x: \varphi(x)=1\}$. 若对某些 x 有 $0<\varphi(x)<1$，则 φ 称为**随机检验**. 因为，为实施这检验，除取得样本 x 外，有时还有必要再作一个随机试验.

随机检验在实际中不常用，但在理论上有一定的重要性. 即使在实用问题中，随机检验有时也有其用处，如上文所述例 3.1 的情况.

(三) 两类错误与功效函数

在检验假设时，可能会发生所作决定与真实情况不符而产生错误. 错误有两类：一是 H 真确但被否定了，称为**第一类错误**（弃真）；二是 H 不真但被接受了，称为**第二类错误**（采伪）. 如在例 3.1，第一类错误是拒收合格批，第二类错误是接受不合格批.

是否犯某一类错误，犯错误可能性大小，取决于参数 θ 的值，

所用的检验和所得到的样本 x. 我们不能要求一个检验方法永远不会出错，但可以要求尽可能使犯错误的概率小一些.

为描述这个问题,要引进下面的重要概念.

定义 3.2 设 φ 是(3.32)的一个检验函数,则

$$\beta_\varphi(\theta)=P_\theta(\text{用检验}\,\phi\,\text{否定了}\,H)=E_\theta\varphi(X),\ \theta\in\Theta \tag{3.35}$$

称为 φ 的**功效函数**，也有称**效函数**、**势函数**的. 若 φ 为非随机检验,其否定域为 $\mathscr{X}_2$,则

$$\beta_\varphi(\theta)=P_\theta(X\in\mathscr{X}_2) \tag{3.36}$$

功效函数就是当样本分布参数等于 θ 时，假设 H 被否定的概率. 就例 3.1 而言,若采用检验

$$\varphi(x)=1\ \text{当}\ x>c,\ \varphi(x)=0\ \text{当}\ x<c,\ \varphi(c)=\gamma, \tag{3.37}$$

则其功效函数为

$$\beta_\varphi(p)=\sum_{i=c+1}^{n}\binom{n}{i}p^i(1-p)^{n-i}+\gamma\binom{n}{c}p^c(1-p)^{n-c},$$
$$0\leqslant p\leqslant 1. \tag{3.38}$$

知道了一个检验 φ 的功效函数 $\beta_\varphi(\theta)$, 就可算出它犯各类错误的概率:

$$\begin{cases}\varphi\ \text{犯第一类错误的概率}=\begin{cases}\beta_\varphi(\theta), & \text{当}\ \theta\in\Theta_H,\\ 0, & \text{当}\ \theta\in\Theta_K;\end{cases}\\ \varphi\ \text{犯第二类错误的概率}=\begin{cases}0, & \text{当}\ \theta\in\Theta_H,\\ 1-\beta_\varphi(\theta), & \text{当}\ \theta\in\Theta_K.\end{cases}\end{cases} \tag{3.39}$$

例如,当 $\theta\in\Theta_H$ 时,H 本来就成立,当然不可能发生采伪(第二类)错误. 应当注意的是: 一个检验犯某一类错误的概率不是一个常数，而是依赖于样本分布的参数.

(四) 检验的水平、真实水平、限定第一类错误概率的原则

设 φ 是(3.32)的一个检验,而 $0\leqslant\alpha\leqslant 1$.

定义 3.3 若 φ 犯第一类错误的概率总不超过 α(不论 θ 在 Θ 内取何值总不超过 α)，则称 α 是检验 φ 的一个水平，而 φ 称为水

平 α 检验.

由(3.39)立知, φ 为水平 α 检验的充要条件是

$$\beta_\varphi(\theta)\leqslant\alpha,\ \text{当}\ \theta\in\Theta_H. \tag{3.40}$$

按这个定义,一个检验的水平不唯一:若 α 是检验 φ 的水平而 $\alpha<\alpha'\leqslant1$, 则 α' 也是检验 φ 的水平. 为免除这种不方便,有时称一检验的最小水平为其真实水平. 显然

$$\varphi\ \text{的真实水平}=\sup\{\beta_\varphi(\theta):\theta\in\Theta\}. \tag{3.41}$$

水平这个概念的产生, 导源于 Neyman-Pearson 提出的一个重要原则——限定第一类错误概率的原则. 按照(3.39), 一个好的检验应使其功效函数在 Θ_H 上取小值, 而在 Θ_K 上取大值. 这一点不能随心所欲地达到: 为使 $\beta_\varphi(\theta)$在 $\theta\in\Theta_H$ 时小, 就要少否定 H, 即要把否定域取得小一些, 而这也就会使 $\beta_\varphi(\theta)$在 Θ_K 上跟着也小了. 反过来也一样. 为解决这个困难, Neyman-Pearson 提出: 指定 α, $0<\alpha<1$. 限制所用的检验有水平 α, 即其第一类错误概率总不超过 α. 在这个限制之下谋求使第二类错误概率尽可能小. 这样, Neyman-Pearson 就得以把假设检验问题提成一个明确的数学最优化问题. 在约束条件"$\beta_\varphi(\theta)\leqslant\alpha$ 当 $\theta\in\Theta_H$" 之下, 使 $\beta_\varphi(\theta)$在 Θ_K 上尽可能大". 怎么叫做"尽可能大"? 这句话可赋予种种意义, 而这就引出一些最优化准则. 怎样制定各种合理的准则,并在这些准则之下找出最优的检验,就成为 NP 理论的中心内容.

至于水平 α 的选择,习惯上一般是把 α 取得比较小且标准化. 例如 $\alpha=0.001$, 0.005, 0.01, 0.05, 0.1 等值, 而不取 0.0412 这种值. 标准化是为了造表方便. 至于 α 习惯上取小值, 是因为在应用上有这样一类问题,其中原假设代表一种久已存在的状态(如一种已用了多时的生产方法), 而对立假设则反映一种改变(如一种刚提出而未经实践考验过的新生产方法),把 α 取得很小使弃真错误概率很小. 因此,一旦检验的结果是否定原假设(即改用新生产方法), 那一定是有充分的根据. 这反映了应用者这样一种心理: 除非证据很有说服力,不轻易改变现有状态而投入那种未经考

验过的、其后果难于估计的新状态中去. 当然，在一具体问题中，当两类错误的后果已有了明确估计时，α 究竟取多大为好，可根据这一具体情况去决定，而不一定拘泥于取小的 α 值.

以上所述就是 NP 理论的要旨. 我们特别强调的是，这个理论明确和新提出了一系列的严格数学概念，尤其是把检验问题提成了一个数学最优化的问题. 第五章将介绍的 Wald 判决函数理论又沿着这个方向作了发展. 所以有人认为，NP 理论是 Wald 理论的先声.

拿 NP 理论与 Fisher 的工作比较，可以看出，诸如原假设、否定域、第一类错误和水平等概念，在 Fisher 的工作中也提出或隐含了. 在例 3.2、例 3.3 和 K. Pearson χ^2 检验中所提到的"阈值"α，实际上就是 NP 理论中检验的水平. 他们用这个阈值 α 保证了：当使用该检验时，原假设被错误地否定的机会只有 α. 但这样的检验很多，用什么标准比较其优劣？Fisher 由于未明确提出对立假设这个概念，也就不可能提出第二种错误概率，而使他的工作中缺少了这个关键之点. NP 理论补足了这一点. 至此可以看出，如我们在前面曾提到的，尽管对立假设这个概念看来一目了然，它的明确提出有重大意义.

§3.4 一致最优检验与无偏检验

(一)一致最优检验 Neyman-Pearson 基本引理

设有检验问题(3.32)，指定 α, $0<\alpha<1$, 以 Φ_α 记(3.32)的一切水平 α 检验的集.

定义 3.4 若 $\varphi\in\Phi_\alpha$，且对任何 $\varphi_1\in\Phi_\alpha$ 有

$$\beta_\varphi(\theta)\geqslant\beta_{\varphi_1}(\theta), \text{当 } \theta\in\Theta_K, \tag{3.42}$$

则称 φ 为(3.32)的一个水平 α 的一致最优检验，简称水平 α 的 **UMP** 检验(UMP 是 Uniformly Most Powerful 的缩写).

按(3.39)，当 φ 为水平 α 的 UMP 检验时，它在限制第一类错误概率不超过 α 的条件下，总使(即对一切 $\theta\in\Theta_K$)第二类错误概

率达到最小. 因此,若以错误概率为衡量检验优劣的唯一尺度,且接受限制第一类错误概率的原则,则 UMP 检验是最好的检验. 不过, UMP 检验的存在一般是例外而不是常见. 理由如下: 若 Θ_K 不止包含一点(这时它称为复合的. 若只含一个点则称为简单的). 则当在其中取两个不同点 θ_1 和 θ_2 时, 为使 $\beta_\varphi(\theta_1)$尽可能大的那种检验 φ,不见得同时也能使 $\beta_\varphi(\theta_2)$大. 如在 §3.2(三)独立性检验中, 对立假设可以是正相关也可以是负相关. 一个针对前者的检验,对后者的功效就很低. 在这种情况下, UMP 检验不可能存在.

在 Θ_K 只包含一个点时, 一般来说 UMP 检验存在. 但确定这个 UMP 检验也不易, 只有在 Θ_H 和 Θ_K 都只包含一个点时, 问题很简单:

定理 3.3(NP 基本引理) 设样本 X 的分布有概率函数 $f(x, \theta)$, 参数 θ 只有两个可能值 θ_0、θ_1, 则对任给 $\alpha\in(0, 1)$存在常数 C 和 $\gamma\in[0, 1]$,使由下式决定的检验函数 φ_α 是检验问题

$$H:\theta=\theta_0\leftrightarrow K:\theta=\theta_1 \tag{3.43}$$

的水平 α 的 UMP 检验:

$$\varphi_\alpha(x)=\begin{cases}1, & >C;\\ \gamma, \quad 当\ f(x, \theta_1)/f(x, \theta_0) & =C;\\ 0, & <C.\end{cases} \tag{3.44}$$

证 以 h 记当 $\theta=\theta_0$ 时, 变量 $f(X, \theta_1)/f(X, \theta_0)$的分布函数. 由于 $0<\alpha<1$, 存在 C, 使 $h(C)\leqslant 1-\alpha\leqslant h(C+0)$(我们采用左连续的分布). 若 $h(C)=1-\alpha$, 取 $\gamma=1$; 若 $h(C+0)=1-\alpha$, 取 $\gamma=0$; 若 $h(C)<1-\alpha<h(C+0)$, 则取 $\gamma=[h(C+0)-(1-\alpha)]/[h(C+0)-h(C)]$. 显然,这样选择的 C 和 γ 满足条件

$$E_{\theta_0}\varphi_\alpha(X)=1-h(C+0)+\gamma[h(C+0)-h(C)]=\alpha, \tag{3.45}$$

即 φ_α 确有水平 α. 现设 φ 为(3.43)的任一水平 α 检验, 并为确定计设 $f(x, \theta)$是样本 X 的密度. 定义样本空间 $\mathscr{X}$ 的子集 $S^+=\{x: \varphi_\alpha(x)>\varphi(x)\}$, $S^-=\{x:\varphi_\alpha(x)<\varphi(x)\}$. 在 $x\in S^+$ 时 $\varphi_\alpha(x)>0$,

由(3.44)知$f(x,\theta_1)/f(x,\theta_0)\geqslant C$. 在$x\in S^-$时有$\varphi_\alpha(x)<1$, 由(3.44)知$f(x,\theta_1)/f(x,\theta_0)\leqslant C$. 由此可知,在整个样本空间上有$[\varphi_\alpha(x)-\varphi(x)][f(x,\theta_1)-Cf(x,\theta_0)]\geqslant 0$,因此

$$\int_{\mathscr{X}}[\varphi_\alpha(x)-\varphi(x)][f(x,\theta_1)-Cf(x,\theta_0)]dx\geqslant 0,$$

即

$$\int_{\mathscr{X}}\varphi_\alpha(x)f(x,\theta_1)dx-\int_{\mathscr{X}}\varphi(x)f(x,\theta_1)dx$$
$$\geqslant C\Big[\int_{\mathscr{X}}\varphi_\alpha(x)f(x,\theta_0)dx-\int_{\mathscr{X}}\varphi(x)f(x,\theta_0)dx\Big]. \tag{3.46}$$

易见$C\geqslant 0$. 事实上, 若$C<0$, 则因概率函数不取负值, 由(3.44)知$\varphi_\alpha\equiv 1$于$\mathscr{X}$上, 这与φ_α有水平$\alpha<1$矛盾. 其次, 由(3.45)得$\int_{\mathscr{X}}\varphi_\alpha(x)f(x,\theta_0)dx=\alpha$, 又因$\varphi$有水平$\alpha$, 知$\int_{\mathscr{X}}\varphi(x)f(x,\theta_0)\,dx\leqslant\alpha$. 由此知(3.46)式右边非负,故

$$\beta_{\varphi_\alpha}(\theta_1)=\int_{\mathscr{X}}\varphi_\alpha(x)f(x,\theta_1)dx\geqslant\int_{\mathscr{X}}\varphi(x)f(x,\theta_1)dx$$
$$=\beta_\varphi(\theta_1).$$

这证明了φ_α为水平α的UMP检验. 定理证毕.

从"似然性"的观点去看NP基本引理是很清楚的: 对每个样本x, θ_1和θ_0的"似然度"分别为$f(x,\theta_1)$和$f(x,\theta_0)$. 比值$f(x,\theta_1)/f(x,\theta_0)$愈大, 就反映在得到样本$x$时, θ愈像是θ_1而非θ_0, 这样的x就愈应倾向于否定$\theta=\theta_0$的假设.

从NP基本引理还看出随机检验的作用. 若局限于非随机检验, 则对某些α, 我们可能无法找到一个检验, 其真实水平恰为α(即: 不一定存在C, 使$h(C)=\alpha$或$h(C+0)=\alpha$). 这时水平α的UMP(非随机)检验就可能不存在.

(二) NP引理用于求UMP检验

Θ_H和Θ_K都只包含一点的这种情况在应用上少见, 故作为一

个求 UMP 检验的方法，NP 基本引理的直接意义不大．但是，通过这个基本引理，可以求得某些常见的检验问题的 UMP 检验，其中 Θ_H 和 Θ_K 不止包含一个点．

在 Θ_H 中挑出一个点 θ_0 尽量接近 Θ_K．在 Θ_K 中任取一点 θ_1，构成检验问题(3.43)．按 NP 基本引理，可求得这问题的 UMP 检验 $\varphi_{\alpha,\theta_1}$．一般，当 θ_1 在 Θ_K 内变化时，$\varphi_{\alpha,\theta_1}$ 也会随之而变．现假定 $\varphi_{\alpha,\theta_1}$ 不依赖于 θ_1，即不论 θ_1 在 Θ_K 内如何变化，$\varphi_{\alpha,\theta_1}$ 总等于一个固定的检验函数 φ_α．若再假定 φ_α 作为(3.32)的检验有水平 α，则它就是(3.32)的水平 α 的 UMP 检验．

事实上，任取(3.32)的一个水平 α 检验 φ，并设 $\theta_1\in\Theta_K$．因 φ 为(3.32)的水平 α 检验，自然有 $\beta_\varphi(\theta_0)\leqslant\alpha$．因此 φ 也是(3.43)的水平 α 检验．但 φ_α 是(3.43)的水平 α 的 UMP 检验，故 $\beta_{\varphi_\alpha}(\theta_1)\geqslant\beta_\varphi(\theta_1)$．由于 θ_1 是在 Θ_K 内任取的，证明了所要的结果．

当然，此法要行得通并不容易，即使(3.32)的 UMP 检验果真存在，也不一定能找到具有上述性质的 θ_0．只有在参数空间 Θ 是一维的(R^1 或其一区间)，而假设是单边的：

$$H:\theta\leqslant\theta_0\leftrightarrow K:\theta>\theta_0, \tag{3.47}$$

且样本 X 的分布为指数型，即其概率函数为

$$f(x,\theta)=C(\theta)e^{Q(\theta)T(x)}h(x). \tag{3.48}$$

其中 $C(\theta)$、$Q(\theta)$ 只是 θ 的函数，$T(x)$、$h(x)$ 只是样本 x 的函数，且 $Q(\theta)$ 为 θ 的严格增加函数之时，UMP检验必存在．这就是下面的重要定理：

定理 3.4 设样本 X 的分布为指数型(3.48)，Θ 为$(-\infty,\infty)$的一有限或无限区间，θ_0 为 Θ 的一个内点，且 $Q(\theta)$ 为 θ 的严增函数，则检验问题(3.47)的水平 α 的 UMP 检验存在$(0<\alpha<1)$，且有形式：

$$\varphi_\alpha(x)=\begin{cases}1, & >C;\\ \gamma, \text{ 当 } T(x) & =C;\\ 0, & <C.\end{cases} \tag{3.49}$$

其中常数 C 和 $\gamma(0\leqslant\gamma\leqslant1)$ 满足条件

$$P_{\theta_0}(T(X)>C)+\gamma P_{\theta_0}(T(X)=C)=\alpha. \qquad (3.50)$$

证 任取 $\theta_1>\theta_0$，构成检验问题(3.43)，有

$$f(x,\theta_1)/f(x,\theta_0)=\frac{C(\theta_1)}{C(\theta_0)}\exp[T(x)(Q(\theta_1)-Q(\theta_0))].$$

由于 $Q(\theta_1)-Q(\theta_0)>0$，且 $C(\theta_1)/C(\theta_0)>0$，上式右边为 $T(x)$ 的严增函数. 因此，比值 $f(x,\theta_1)/f(x,\theta_0)$ 大于、等于或小于某一常数 C'，分别相应于 $T(x)$ 大于、等于或小于某一常数 C. 由此及 NP 基本引理，知(3.43)的水平 α 的 UMP 检验有形式(3.49)，其中 C 和 γ 由(3.50)确定. 由于 C 和 γ 与 θ_1 无关，根据前面的讨论知，为了证明由(3.49)和(3.50)确定的检验 φ 是(3.47)的水平 α 的 UMP 检验，只须证明：φ 是(3.47)的水平 α 检验. 显然，欲证明这一点，只须证明：φ 的功效函数 $\beta_\varphi(\theta)$ 是 θ 的非降函数. 下面来证明这一点.

任取 $\theta'<\theta''$，由(3.48)知

$$f(x,\theta')/f(x,\theta'')=\frac{C(\theta')}{C(\theta'')}\exp[(Q(\theta')-Q(\theta''))T(x)].$$

因为 Q 为 θ 的严格增加函数，且 $\theta'<\theta''$，有 $Q(\theta')-Q(\theta'')<0$. 又 $C(\theta')>0$, $C(\theta'')>0$. 由此知 $f(x,\theta')/f(x,\theta'')$ 只与 $T(x)$ 有关，且是 $T(x)$ 的严格下降函数. 找 t_0，使

$$\frac{C(\theta')}{C(\theta'')}\exp[(Q(\theta')-Q(\theta''))t_0]=1.$$

这样的 t_0 必存在，否则恒有 $f(x,\theta')<f(x,\theta'')$ 或 $f(x,\theta')>f(x,\theta'')$，这与 $\int_{\mathscr{X}}f(x,\theta')dx=\int_{\mathscr{X}}f(x,\theta'')dx=1$ 矛盾. 令

$$S_1=\{x:T(x)>t_0\},\ S_2=\{x:T(x)<t_0\},$$
$$S_3=\{x:T(x)=t_0\},$$

则由 $\int_{\mathscr{X}}f(x,\theta')dx=\int_{\mathscr{X}}f(x,\theta'')dx=1$，以及 t_0 的定义，易知

$$0\leqslant\int_{S_1}[f(x,\theta'')-f(x,\theta')]dx$$
$$=-\int_{S_2}[f(x,\theta'')-f(x,\theta')]dx.$$

由(3.49)知，$\varphi(x)$只与$T(x)$有关，且是$T(x)$的非降函数，故有$\inf\limits_{x\in S_1}\varphi(x)\geqslant\sup\limits_{x\in S_2}\varphi(x)$. 因此

$$
\begin{aligned}
\beta_\varphi(\theta'')-\beta_\varphi(\theta') &= \int_{\mathscr{X}}\varphi(x)[f(x,\theta'')-f(x,\theta')]\,dx \\
&= \int_{S_1}\varphi(x)[f(x,\theta'')-f(x,\theta')]\,dx \\
&\quad + \int_{S_2}\varphi(x)[f(x,\theta'')-f(x,\theta')]\,dx \\
&\geqslant [\inf_{x\in S_1}\varphi(x)-\sup_{x\in S_2}\varphi(x)] \\
&\quad \times\int_{S_1}[f(x,\theta'')-f(x,\theta')]\,dx\geqslant 0.
\end{aligned}
$$

这证明了$\beta_\varphi(\theta')\leqslant\beta_\varphi(\theta'')$，即$\beta_\varphi(\theta)$非降，因而完成了定理的证明.

完全同样的方法证明：若考虑检验问题

$$H:\theta\geqslant\theta_0\leftrightarrow K:\theta<\theta_0,$$

则对任给的$\alpha\in(0,1)$，水平α的UMP检验存在，而且有形式

$$
\varphi_\alpha^*(x)=\begin{cases}1, & <C;\\ \gamma, \quad \text{当 } T(x) & =C;\\ 0, & >C.\end{cases} \tag{3.51}
$$

其中常数C和$\gamma(0\leqslant\gamma\leqslant1)$满足条件

$$P_{\theta_0}(T(X)<C)+\gamma P_{\theta_0}(T(X)=C)=\alpha. \tag{3.52}$$

例 3.9 考虑例3.1. 此处$\mathscr{X}=\{0,1,\cdots,n\}$，$\Theta=\{p:0<p<1\}$. 有$f(x,p)=\binom{n}{x}p^x(1-p)^{n-x}$，可写为$(1-p)^n\exp\left[x\times\log\frac{p}{1-p}\right]$，$Q(p)=\log\frac{p}{1-p}$在$0<p<1$时严增，$T(x)=x$. 故定理3.4的条件全满足，因而问题$H:p\leqslant p'\leftrightarrow K:p>p'$的水平$\alpha$的UMP检验$\varphi$存在，且有(3.37)的形式，其中$C$和$\gamma$由(3.38)决定.

本例中参数空间可取为$\{p:0\leqslant p\leqslant1\}$. 例中为了写成指数型，

舍弃掉了0、1两个极端值，不难证明：即使保留这两个值，例中的结果仍成立，这留给读者作为一个练习。

例3.10 设 $X_1, \cdots, X_n \sim N(\theta, \sigma_0^2)$，$\sigma_0$ 已知．要检验假设 $H:\theta\leqslant 0 \leftrightarrow K:\theta>0$．有

$$f(x_1, \cdots, x_n, \theta)=(2\pi\sigma_0^2)^{-n/2}\exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(x_i-\theta)^2\right]$$

$$=(2\pi\sigma_0^2)^{-n/2}e^{-n\theta^2/2\sigma_0^2}\exp\left[\frac{n}{\sigma_0^2}\theta\bar{x}\right]$$

$$\times\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n x_i^2\right). \tag{3.53}$$

这是(3.48)的形状，其中 $Q(\theta)=n\theta/\sigma_0^2$ 严增，$T(x)=\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i$ $(x=(x_1, \cdots, x_n))$．据定理3.4，知水平 α 的UMP检验存在，且有否定域 $\{\bar{x}>C\}$，常数 C 由 $P_0(\overline{X}>C)=\alpha$ 决定，易见 $C=\sigma_0 u_\alpha/\sqrt{n}$，其中 u_α 由条件

$$\frac{1}{\sqrt{2\pi}}\int_{u_\alpha}^{\infty}e^{-t^2/2}dt=\alpha \tag{3.54}$$

决定，可由正态分布表查出．本例所以UMP检验为非随机的，是因为 $\overline{X}$ 的分布为连续型，它取单个值的概率为0．在连续型分布中多是这种情况．

另有几个可表为单参数指数型分布族的重要情况，留作为习题．也有个别非指数型的情况，UMP检验也存在．一个重要例子是习题2．至于在多参数的情况，UMP检验的存在就是很稀有的例外了．有一个重要例子是美国统计学家E. L. Lehmann和C. Stein在1948年得到的：设 $X_1, \cdots, X_n \sim N(a, \sigma^2)$，$a$、$\sigma$ 都未知(两参数)，要检验

$$H:\sigma\leqslant\sigma_0,\ a\text{任意}\leftrightarrow K:\sigma>\sigma_0,\ a\text{任意}. \tag{3.55}$$

他们证明了：(3.55)的水平 α 的UMP检验存在，且有否定域 $\{(x_1, \cdots, x_n):\sum_{i=1}^n(x_i-\overline{X})^2\geqslant\sigma_0^2\chi_{n-1}^2(\alpha)\}$．但这个结果不能用本节的方法证明．

(三)无偏检验　一致最优无偏检验

前已说明,UMP 检验的存在是少有的例外. 因此, 作为一个准则,它的实际效用是很有限的. 为了得到适用范围更大的准则,有两个途径可以采取. 其一是引进一种综合性的指标. 就是说,对(3.32)的每一个检验 φ, 定义一个数量指标 $M(\varphi)$. 若 φ_1, φ_2 是(3.32)的两个检验,则当 $M(\varphi_1)>M(\varphi_2)$时, 称 φ_1 优于 φ_2. 若 $M(\varphi)\geqslant M(\varphi_1)$对任一检验 φ_1, 则称 φ 在准则M之下为最优的. 也可以先指定一个 α, 在一切水平为 α 的检验中, 去寻找按准则M 为最优的检验. 另一个途径是先对所考虑的检验施加某种合理的、一般性的限制,这样就缩小了所考虑的检验的范围,然后在这缩小了的范围内去寻找一致最优的检验. 正如在点估计中, 我们先限制估计量必须是无偏的,然后在无偏估计类中,去寻找在某一准则下最优的估计,例如 UMVUE.

这后一途径的一个重要例子是**无偏检验**. 我们先给它下一个正式的定义.

定义 3.5　设 φ 为(3.32)的一个检验, $\beta_\varphi(\theta)$为其功效函数. 若对任何 $\theta_1\in\Theta_H$ 及 $\theta_2\in\Theta_K$, 总有 $\beta_\varphi(\theta_1)\leqslant\beta_\varphi(\theta_2)$, 则称 φ 是(3.32)的一个无偏检验. 若无偏检验 φ 有水平 α, 则称 φ 为水平 α 的无偏检验.

无偏检验的直观意义很清楚: 若 φ 为 $H\leftrightarrow K$ 的无偏检验, 则当 H 正确时, H 被否定(这否定是错的)的概率, 不应超过当 H 不正确时, H 被否定(这否定是对的)的概率. "无偏"一词还可以这样去理解: 这个检验考虑到了对立假设 Θ_K 中各种参数值, 不因特别"照顾"某一部分参数值而"牺牲"另一部分参数值. 如下面的例子.

例 3.11　$X_1, \cdots, X_n\sim N(\theta, 1)$, 检验问题为

$$H:\theta=0\leftrightarrow K:\theta\neq 0. \tag{3.56}$$

取水平 α. 若特别照顾 $\theta>0$ 这部分参数值, 则取以$\{\bar{X}>u_\alpha/\sqrt{n}\}$为否定域的检验最好. 但这个检验对 $\theta<0$ 的这部分对立假设参

数值表现很差：其功效在这些点处小于水平 α，且当 $\theta\to-\infty$ 时趋于 0. 同样，若特别照顾 $\theta<0$ 这部分参数值，则取以 $\{\overline{X}<-u_\alpha/\sqrt{n}\}$ 为否定域的检验最好. 但它在对立假设 $\theta>0$ 处表现很差. 这两个检验都不是无偏的. 容易证明(习题 9)：以 $\{|\overline{X}|>u_{\alpha/2}/\sqrt{n}\}$ 为否定域的检验是无偏检验. 这个检验考虑了 $\theta=0$ 两边的参数值.

任一检验问题都存在各水平的无偏检验. 比方说，$\varphi(x)\equiv\alpha$ 就是其一，因为其功效函数在 Θ 上处处等于 α. 以 U_α 记(3.32)的一切水平 α 无偏估计的类.

定义 3.6 若 $\varphi\in U_\alpha$，且对任何 $\varphi_1\in U_\alpha$，有

$$\beta_\varphi(\theta)\geqslant\beta_{\varphi_1}(\theta),\ 对一切\ \theta\in\Theta_K, \tag{3.57}$$

则称 φ 是(3.32)的一个水平 α 的**一致最优无偏检验**（简称 UMPU 检验）.

UMPU 检验存在的情况，比 UMP 检验存在的情况要广一些，但仍不是很广. 大体上说，只有在样本 X 的概率函数有指数形状($\theta=(\theta_1,\ \cdots,\ \theta_k)$)

$$f(x,\ \theta)=C(\theta)\exp\left(\sum_{i=1}^{k}\theta_iT_i(x)\right)h(x). \tag{3.58}$$

而假设 H 只涉及其中一个参数(比方说 θ_1). 或诸参数的一个线性型时，UMPU 检验才可能存在. 更确切地说，原假设 H 必须有以下几种形式之一(记 $l=a_1\theta_1+\cdots+a_k\theta_k$，$a_1,\ \cdots,\ a_k$ 为已知常数)：

$$H_1: l=l_0;\ H_2: l\leqslant l_0;\ H_3: l\geqslant l_0;\ H_4: l_1\leqslant l\leqslant l_2. \tag{3.59}$$

有时，X 的概率函数不直接以(3.58)的形式出现，但可以对参数 θ 作一定的变换达到这一点. 例如，$X_1,\ \cdots,\ X_n\sim N(a,\ \sigma^2)$，概率函数为

$$f(x_1,\ \cdots,\ x_n,\ a,\ \sigma)=(\sqrt{2\pi}\,\sigma)^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right)$$

$$=(\sqrt{2\pi}\ \sigma)^{-n}e^{-na^2/2\sigma^2}$$

$$\times \exp\left[\frac{a}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right]. \qquad (3.60)$$

以 $\theta_1=\frac{a}{\sigma^2}$, $\theta_2=-\frac{1}{2\sigma^2}$ 为新参数,并记 $T_1(x)=\sum_1^n x_i$, $T_2(x)=\sum_1^n x_i^2$, 则(3.60)可写为

$$f(x_1, \cdots, x_n, \theta_1, \theta_2)=\left(-\frac{\theta_2}{\pi}\right)^{n/2} e^{n\theta_1^2/4\theta_2}$$
$$\times \exp(\theta_1 T_1(x)+\theta_2 T_2(x)). \qquad (3.61)$$

其中 θ_1, θ_2 变化范围为 $-\infty<\theta_1<\infty$, $\theta_2<0$. 有关 a 和 σ 的一些检验问题,可化为(3.59)的形状,如:

$$a=a_0 \rightarrow \theta_1-2a_0\theta_2=0;\ a\leqslant a_0 \rightarrow \theta_1-2a_0\theta_2\leqslant 0;$$
$$\sigma^2=\sigma_0^2\rightarrow \theta_2=-\frac{1}{2\sigma_0^2};\quad \sigma^2\leqslant\sigma_0^2 \rightarrow \theta_2\leqslant -\frac{1}{2\sigma_0^2}.$$

这些假设都可以找到 UMPU 检验,有关定理的确切表述及其严格数学证明很复杂,超出了基础课的范围之外,故在此都从略了. 有兴趣的读者可参看陈希孺《数理统计引论》§ 3.3.

§ 3.5 似然比检验

似然比检验是 Neyman 和 Pearson 提出的一种构造检验的方法. 这是一种基于直观想法的方法,有如点估计中的极大似然估计. 用这种方法构造出来的检验,一般说有比较良好的性质,但是并不能一般地证明,这样构造出的检验必然满足某些常见的最优准则,如 UMP 或 UMPU之类. 这个方法的一个优点是适用面较广,就是说,它对分布族的形式没有什么特殊的要求.

(一)似然比检验的定义

定义 3.7 设样本 X 有概率函数 $f(x, \theta)$, $\theta\in\Theta$, Θ_H 为 Θ 的非空真子集. 考虑假设检验问题(3.32),则统计量

$$LR(x)=\sup_{\theta\in\Theta} f(x, \theta)/\sup_{\theta\in\Theta_H} f(x, \theta) \qquad (3.62)$$

称为关于该检验问题的似然比. 而由下式定义的检验函数 φ:

$$\varphi(x)=\begin{cases}1, & >C;\\ \gamma, & \text{当 } LR(x)=C;\\ 0, & <C,\end{cases} \tag{3.63}$$

其中 C, γ 为常数, $0\leqslant\gamma\leqslant1$, 称为(3.32)的一个似然比检验.

当 Θ_H 和 Θ_K 都只包含一个点时, 似然比检验(3.63)就是 NP 基本引理中确定的那种检验(3.44)(严格讲, 这句话只在一定条件下才正确). 在 §3.4(一)的末尾处, 我们曾对检验(3.44)的直观根据作过说明. 该说明大体上也适用于似然比检验: 有了样本 x 后, 似然性最大的参数值, 是 θ 的极大似然估计 $\hat{\theta}$. $\hat{\theta}$ 一般与参数真值 θ_0 应接近. (3.62)右边的分子当然等于 $f(x, \hat{\theta})$. 若假设 H 成立, 即 θ_0 确在 Θ_H 内, 则(3.62)的分母 $\geqslant f(x, \theta_0)\approx f(x, \hat{\theta})$, 这时 $LR(x)$ 接近于 1; 反过来, 若 H 不成立, 则 θ_0 不在 Θ_H 内, 因 $\hat{\theta}$ 与 θ_0 接近, $\hat{\theta}$ 也不在 Θ_H 内. 这时(3.62)的分母将小于 $f(x, \hat{\theta})$, 而 $LR(x)$ 将比较显著地大于 1. 因此反过来说, $LR(x)$ 取大值可以看作 $\theta_0\bar{\in}\Theta_H$, 即 H 不正确的证据. 举两个简单例子.

例 3.12 $X_1, \cdots, X_n\sim N(\theta, 1)$, $H:\theta=\theta_0\leftrightarrow K:\theta\neq\theta_0$, θ_0 给定, 有 $f(x_1, \cdots, x_n, \theta)=(2\pi)^{-n/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right]$. 显见 $\sup\limits_{\theta\in\Theta_H}f(x_1,\cdots, x_n, \theta)=f(x_1, \cdots, x_n, \theta_0)=(2\pi)^{-n/2}\exp\left[-\frac{1}{2}\times\sum_{i=1}^{n}(x_i-\theta_0)^2\right]$, $\sup\limits_{\theta\in\Theta}f(x_1, \cdots, x_n, \theta)=(2\pi)^{-n/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$. 于是

$$LR(x_1, \cdots, x_n)=\exp\left[\frac{1}{2}n(\bar{x}-\theta_0)^2\right],$$

$LR(x_1, \cdots, x_n)$是$|\bar{x}-\theta_0|$的严格增加函数. 由此导出似然比检验有否定域$\{|\bar{x}-\theta_0|>C\}$($\overline{X}$ 的分布为连续型, (3.63) 中取 γ 那一栏没有必要). C 之值要根据检验的水平来定: 若取水平 α, 则应取 $C=u_\alpha/\sqrt{n}$.

如果原假设为 $\theta\leqslant\theta_0$ 而对立假设为 $\theta>0$, 则不难算得

$$LR(x_1, \cdots, x_n)=\begin{cases}\exp\left[\frac{1}{2}n(\bar{x}-\theta_0)^2\right], & 当\ \bar{x}>\theta_0;\\ 1, & 当\ \bar{x}\leqslant\theta_0.\end{cases}$$

(3.64)

这时，若取水平 $\alpha\leqslant 1/2$(这是常见的情况)，则似然比检验将有否定域 $\{\bar{x}-\theta_0>u_\alpha/\sqrt{n}\}$. 这个简单推理留给读者（习题）. 但若取水平 $\alpha>1/2$ 而坚持要用似然比检验，则必须用如下的检验 φ:

$$\varphi(x_1, \cdots, x_n)=\begin{cases}1, & 当\ \bar{x}>\theta_0;\\ 2\alpha-1, & 当\ \bar{x}\leqslant\theta_0.\end{cases} \tag{3.65}$$

(习题 13)在前一情况 $\left(\alpha\leqslant\frac{1}{2}\right)$, 似然比检验为 UMP, 而在后一情况则否. 在后一情况其实 UMP 检验仍存在(例 3.10). 由此可见，似然比检验不必给出最优的检验.

问题的症结在于: 在原假设为 $\theta=\theta_0$ 时, Θ_H 为零维集而 Θ 为一维集，维数不等; 在原假设为 $\theta\leqslant\theta_0$ 时, Θ_H 和 Θ 同为一维集. 只要 Θ_H 与 Θ 的维数相同，本例这样的情况在水平 α 较大时就可能出现. 不过这个现象并非很重要: 因为，一则在应用上水平 α 总取得较小，这时上述情况不大发生. 二则即使 α 较大而上述情况发生了，我们也可以参照 α 较小时的似然比检验去修正之(实际上，就是只取似然比中不等于 1 的那一部分，把它看作是在整个样本空间上都成立). 在本节(三)中我们还会看到一些这样的例子.

例 3.13 回到 §3.2(一)中K. Pearson χ^2拟合优度检验问题那个最简单的情况. 这相当于给定了 $p_1, \cdots, p_r$, 都大于 0, 其和为 1. 又有了样本$(\nu_1, \cdots, \nu_r)$, 其概率分布为 $P_{\theta_1\cdots\theta_r}(\nu_1, \cdots, \nu_r)=\frac{n!}{\nu_1!\cdots\nu_r!}\theta_1^{\nu_1}\cdots\theta_r^{\nu_r}$, 要检验假设 $H:\theta_1=p_1, \cdots, \theta_r=p_r$. 此处 $\Theta=\{(\theta_1, \cdots, \theta_r):\theta_i\geqslant 0,\ i=1, \cdots, r, \sum_1^r\theta_i=1\}$, 而 Θ_H 只包含一点$(p_1, \cdots, p_r)$. 简单计算给出

$$LR(\nu_1, \cdots, \nu_r)=\prod_{i=1}^r\left(\frac{\nu_i}{np_i}\right)^{\nu_i}. \tag{3.66}$$

与(3.9)比较，可见此问题的似然比检验不同于 K. Pearson 的 χ^2

检验．但在样本大小 $n\to\infty$ 时，在下一段中我们将指出，二者渐近地等价．

（二）似然比的渐近分布

为了定出(3.63)中的 C 和 γ，使它有给定的水平 α，就需要知道似然比 $LR(X)$ 在原假设成立时的分布．在简单例子中，似然比的分布可以算出．但在多数情况下，如在例3.13，似然比有很复杂的形状，其精确分布无法求得．1938年，S. S. Wilks 证明了：若样本 X 是独立随机样本 $X_1, \cdots, X_n$，则当样本大小 $n\to\infty$ 时，在原假设成立之下，似然比有一个简单的极限分布．应用这极限分布，可以近似地决定(3.63)中的 C 和 γ．

Wilks 定理的确切陈述包含一大堆关于总体的概率函数的假定，需要两页纸才能写下来，其证明也很复杂．因此，我们不打算作完整的叙述，只指明条件中的一个至关重要之点，就是 Θ 的维数应当高于 Θ_H 的维数．举例而言，若样本 $X_1, \cdots, X_n \sim N(a, \sigma^2)$，原假设为 H：$a=a_0$，则 Θ 是 R^2 中的上半平面$\{(a, \sigma): -\infty<a<\infty, \sigma>0\}$．这是一个二维集．$\Theta_H$ 是 Θ 内的一条半直线：$\Theta_H=\{(a, \sigma): a=a_0, \sigma>0\}$．虽然它身居平面上，但是只是一个一维集．又如，三维空间中的球面是二维集，球体是三维集．一个点是零维集．在例 3.13 中，参数为 $\theta_1, \cdots, \theta_r$，但其和为 1，故实质上只有参数 $\theta_1, \cdots, \theta_{r-1}$，而 $\Theta=\{(\theta_1, \cdots, \theta_{r-1}): \theta_1\geqslant 0, \cdots, \theta_{r-1}\geqslant 0, \sum\limits_{1}^{r-1}\theta_i\leqslant 1\}$．这是 R^{r-1} 空间中的一个有内点的集，为 $r-1$ 维集．本例的 Θ_H 只含一个点，为零维集．明确了这一点，可以把 Wilks 的定理大致地表述为：

定理 3.5 若 Θ 的维数 $-$ Θ_H 的维数 $=t>0$，则在原假设 $\theta\in\Theta_H$ 成立之下，当样本大小 $n\to\infty$ 时，$2\log LR(X_1, \cdots, X_n)$ 有极限分布 χ_t^2．

定理的确切陈述和证明可参看陈希孺《数理统计引论》p. 326～330．还有一点需要明确：原假设 Θ_H 中可以包含不止一个点．这

时，定理3.5的含义是：不论参数真值 θ 落在 Θ_H 内何处，$2\log LR$ 的极限分布总是自由度为 t 的 χ^2 分布 χ^2_t.

利用这个定理可近似决定(3.63)中的 C 和 γ. 为此，把 $2\log LR(X_1, \cdots, X_n)$ 的分布看成是确切地等于 χ^2_t，则应取

$$C=\exp\left[\frac{1}{2}\chi^2_t(\alpha)\right], \gamma=0. \tag{3.67}$$

当 C 和 γ 按(3.67)式取时，由(3.63)确定的检验 φ 满足

$$\lim_{n\to\infty}\beta_\varphi(\theta)=\alpha, \text{对任何}\theta\in\Theta_H. \tag{3.68}$$

常把满足这条件的检验 φ 称为有**渐近水平** α.

我们就例3.12和3.13来验证一下这定理的正确性. 在例3.12中，$2\log LR(X_1, \cdots, X_n)=n(\overline{X}-\theta_0)^2$. 当原假设 $\theta=\theta_0$ 成立时，$\sqrt{n}(\overline{X}-\theta_0)\sim N(0, 1)$. 故依定义有 $n(\overline{X}-\theta_0)^2\sim\chi^2_1$. 在这个场合，$2\log LR$ 的精确分布就是其极限分布. 本例 $t=1-0=1$.

例3.13较复杂. 记 $\mu_i=\nu_i-np_i$，由(3.66)有

$$\log LR=\sum_{i=1}^r \nu_i\log\frac{\nu_i}{np_i}=\sum_{i=0}^r(np_i+\mu_i)\log\left(1+\frac{\nu_i}{np_i}\right). \tag{3.69}$$

在原假设成立时，按大数定律，有 $\nu_i/n\to p_i$ 当 $n\to\infty$，即 $\mu_i/n\to 0$ (依概率或以概率1). 于是按(3.69)，有

$$\log LR=\sum_{i=1}^r(np_i+\mu_i)\left[\frac{\mu_i}{np_i}-\frac{1}{2}\left(\frac{\mu_i}{np_i}\right)^2+O\left(\frac{\mu_i^3}{n^3}\right)\right].$$

注意到 $\sum_1^r\mu_i=0$，由上式得

$$2\log LR=\sum_{i=1}^r\mu_i^2/np_i+O\left(\sum_{i=1}^r\mu_i^3/n^2\right). \tag{3.70}$$

我们来证明：$\mu_i^3/n^2\xrightarrow{P}0$，若 $n\to\infty$ 且原假设成立. 事实上，依中心极限定理，有 $\mu_i/\sqrt{np_i(1-p_i)}\xrightarrow{\mathscr{L}}N(0, 1)$. 由此知 $\mu_i^3/n^{3/2}$ 当 $n\to\infty$ 时有极限分布. 因为 $\mu_i^3/n^2=\frac{1}{\sqrt{n}}(\mu_i^3/n^{3/2})$ 而 $\frac{1}{\sqrt{n}}\to 0$,

知 $\mu_i^3/n^2 \xrightarrow{p} 0$(在证明定理 2.6 时，已用过这个事实)．于是由(3.70)知，在原假设成立而 $n\to\infty$ 时，$2\log LR$ 与 $\sum_1^r \mu^2/np_i$，即(3.9)式的 k，有同一极限分布．故由定理 3.1，推出在本例中，$2\log LR(X)$ 的极限分布是 χ_{r-1}^2，而 $r-1$ 正是 Θ 的维数与 Θ_H 的维数之差．也可以反过来看：从定理 3.5 出发，把定理 3.1 作为其推论．

利用当样本大小 $n\to\infty$ 时，统计量的极限分布所构造的检验，称为**大样本检验**．由于不知道统计量的精确分布是常有的事，这类大样本检验在应用上有重要的意义．

例如，有两个总体，其均值分别为 θ_1 和 θ_2．要检验假设 H：$\theta_1=\theta_2\leftrightarrow K:\theta_1\neq\theta_2$．在这两个总体中分别抽出独立随机样本 X_1，…，X_{n_1} 和 Y_1，…，Y_{n_2}．记 $\bar{X}=\frac{1}{n_1}\sum_1^{n_1}X_i$，$\bar{Y}=\frac{1}{n_2}\sum_1^{n_2}Y_i$．设两个总体的方差 σ_1^2 和 σ_2^2 都非 0 有限，则根据中心极限定理，当原假设成立即 $\theta_1=\theta_2$ 时，$(\bar{X}-\bar{Y})/\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$ 的分布当 $n_1\to\infty$ 和 $n_2\to\infty$ 时收敛于标准正态分布 $N(0, 1)$．又

$$S_1^2=\sum_1^{n_1}(X_i-\bar{X})^2/(n_1-1) \quad 和 \quad S_2^2=\sum_1^{n_2}(Y_i-\bar{Y})^2/(n_2-1)$$

分别是 σ_1^2 和 σ_2^2 的相合估计．故若令

$$T=(\bar{X}-\bar{Y})/\sqrt{S_1^2/n_1+S_2^2/n_2}, \tag{3.71}$$

则当原假设成立且 $n_1\to\infty$，$n_2\to\infty$ 时，T 的分布也收敛于 $N(0, 1)$．因此，以 $|T|>u_{\alpha/2}$ 为否定域的检验，当 $n_1\to\infty$，$n_2\to\infty$ 时渐近地有水平 α (确切意义如(3.68)式)．

似然比检验的一项最重要的应用，是处理涉及正态分布参数的假设检验问题．这个内容我们将在下节中集中论述．在那里我们将指出，就对正态分布参数的应用而言，似然比检验具有良好的性质．一般地说，用似然比方法作出的检验，在绝大多数情况下表现是不错的，但也有例外情况，见习题 15．

§3.6　正态分布参数的检验及有关检验

(一)引言

在一般情况下,有理由假定,试验的随机误差近似地服从正态分布. 因此,在涉及这种随机误差的大量应用问题中,不少可以归结为正态分布参数的检验,特别是正态分布均值的检验,这就是著名的一样本和两样本 t 检验. 它们都是似然比检验.

然而,在有些情况下,有足够的理由怀疑,随机误差的分布可能与正态分布有较大的偏离. 这时,检验问题的提法就需要作相应的修改,而导致不同的检验方法. 由于这种问题在应用上很重要,我们也在这一节中附带地讨论一下. 系统地处理这种问题属于非参数统计这个专门分支的范围.

(二)一样本 t 检验

设有样本 $X_1, \cdots, X_n \sim N(a, \sigma^2)$, a 和 σ 都是未知参数, 考虑检验问题

$$H: a=a_0 \leftrightarrow K: a \neq a_0, \tag{3.72}$$

a_0 是给定的已知数.

在实际应用中,这个问题最常在以下两种情况下发生:

1. a_0 是一个标准值, 例如, 某种化工产品中所含某物质的量. 由于大批生产中随机性因素的干扰, 以及可能存在的系统性因素的干扰, 产品中所含该物质的量不会总等于 a_0. 随机抽取该产品 n 份作分析, 定出其中该物质含量分别为 $X_1, \cdots, X_n$. 若假定随机误差有正态分布 $N(0, \sigma^2)$, 则可认为 $X_1, \cdots, X_n \sim N(a, \sigma^2)$. 如果 $a=a_0$, 则系统性因素不存在, 产品中所含该物质的量与 a_0 的偏差可全归结于随机因素. 若 $a \neq a_0$, 则表示有某种系统性原因存在,使产品所含该物质的平均量与标准值 a_0 不同. 这时就有必要找出原因,以消除这种系统性因素.

2. 两处理的对比试验. 在数理统计学中,"处理"一词的含义

极广．它可以表示一种工艺流程，一个种子品种，一种治疗方法等等．现有两种处理，为确定计设想为两种施肥方法，要设计一种试验，以判明其效应是否有差别．

选择 n 对试验单元，每对中的两个试验单元条件尽可能均匀，而不同对中的单元则可有较大差别．在每一对单元中，随机地决定其中之一施加处理甲，另一施加处理乙，正如在例 3.3 中所做的那样，而记录其指标值(如单位面积产量)．以 X_i 记第 i 对单元中，甲的指标值与乙的指标值之差，$X_1, \cdots, X_n$ 就是我们的样本．若假定试验的随机误差服从正态分布，则有 $X_1, \cdots, X_n \sim N(a, \sigma^2)$．“甲、乙两处理的效应无差别”相当于 $a=0$．于是我们得到问题(3.72)当 $a_0=0$ 的情形．

现考虑(3.72)的水平 α 似然比检验．样本$(X_1, \cdots, X_n)$的似然函数为

$$f(x_1, \cdots, x_n, a, \sigma)=(\sqrt{2\pi}\sigma)^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right). \tag{3.73}$$

考虑到

$$\min_{-\infty<a<\infty}\sum_{i=1}^{n}(x_i-a)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2, \tag{3.74}$$

$$\max_{0<\sigma<\infty}\sigma^{-n}e^{-A/\sigma^2}=(n/2A)^{n/2}e^{-n/2}.\quad (A>0) \tag{3.75}$$

易算出似然比为

$$\begin{aligned}LR(x_1, \cdots, x_n)&=\left[\sum_{i=1}^{n}(x_i-a_0)^2\Big/\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{n/2}\\&=\left[1+n\left(|\bar{x}-a_0|\Big/\sqrt{\sum_1^n(x_i-\bar{x})^2}\right)^2\right]^{n/2}.\end{aligned}$$

由此知似然比检验有否定域$\left\{|\bar{X}-a_0|\Big/\sqrt{\sum_1^n(X_i-\bar{X})^2}>C\right\}$．为确定常数 C，考虑统计量

$$T=\sqrt{n}(\bar{X}-a_0)\Big/\sqrt{\frac{1}{n-1}\sum_1^n(X_i-\bar{X})^2}. \tag{3.76}$$

由定理 1.1 及 t 分布的定义，知当 $a=a_0$ 时 $T\sim t_{n-1}$，即自由度为 $n-1$ 的 t 分布．给定 $\alpha\in(0, 1)$，可从 t 分布表上，查出满足

$P\left(t_{n-1}>t_{n-1}\left(\frac{\alpha}{2}\right)\right)$的值 $t_{n-1}\left(\frac{\alpha}{2}\right)$. 由于 t 分布密度关于 0 对称，此值也满足 $P\left(|t_{n-1}|>t_{n-1}\left(\frac{\alpha}{2}\right)\right)=\alpha$. 于是得到(3.72)的水平 α 的似然比检验为:

$$\text{当 } |T|>t_{n-1}\left(\frac{\alpha}{2}\right) \text{时否定 } H, \text{不然就接受 } H. \tag{3.77}$$

由(3.76)定义的统计量 T 称为**一样本 t 统计量**，而检验(3.77)则称为**一样本 t 验检**. 统计量 T 在 $a=a_0$ 成立时的密度函数——即在(1.31)式中改 n 为 $n-1$，是英国统计学家 W. S. Gosset首先发现的. 他在 1908 年用"Student"的笔名，在 Biometrika 杂志上发表了这个结果. 因此，t 分布也常称为 Student 分布. 这个结果的发现在数理统计学的发展史上是一件大事，因为它开了小样本理论发展的先声. 在这以前，(3.76)的统计量 T 的分布是近似地作为正态分布来处理. 在此应提到的是: Gosset 本人并未得出 T 的分布的严格证明: 他是在"T 的分布属于 Pearson 分布族"这个假定之下，得出 T 的密度的. 严格的证明是 R. A. Fisher 在几年后作出的. Fisher 与 Gosset 有长期的通信关系. 在这些通信中，Fisher 发展了用 n 维几何(把样本看作 R^n 中的一个点)处理抽样分布的方法，并形成了重要的"自由度"概念.

有时，实际问题归结为检验问题是"单边"形式:

$$H: a\leqslant a_0 \leftrightarrow K: a>a_0. \tag{3.78}$$

如在上述化工产品的例中，$N(a, \sigma^2)$是产品所含某种杂质的量的分布，$a\leqslant a_0$ 表示杂质平均含量不超过某个允许值 a_0. 在两个处理比较的对比试验中，$a\leqslant 0$ 表示"处理甲不优于乙".

注意到当 $\bar{x}>a_0$ 时，有

$$\min_{a\leqslant a_0}\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-a_0)^2.$$

易得(3.78)的似然比为

$$LR(x_1, \cdots, x_n)=\begin{cases}\left[\sum_1^n(x_i-a_0)^2\Big/\sum_1^n(x_i-\bar{x})^2\right]^{n/2}, & \bar{x}>a_0;\\ 1, & \bar{x}\leqslant a_0.\end{cases}$$

与例 3.12 的后一部分相似，并利用刚才得出的结果，易知当 $0<\alpha\leqslant 1/2$ 时，(3.78)的水平 α 似然比检验为：

$$\text{当 } T>t_{n-1}(\alpha) \text{ 时否定 } H\text{，不然就接受 } H. \tag{3.79}$$

(3.79)也称为一样本 t 检验．有时分别称(3.77)为**双边**或**双侧的**，而(3.79)为**单边**或**单侧的**．这名词可从两个意义去理解：(3.72)的对立假设 K 居于原假设 H 的两边，而(3.78)则居于一边．或者，检验(3.77)的否定域两端都有，而(3.79)则只有一端．

若 $\alpha>1/2$，则似然比检验是类似于(3.65)那样的随机检验．此时我们不用似然比检验而仍用检验(3.79)．

可以证明：检验(3.77)和(3.79)分别是(3.72)和(3.78)的水平 α 的 UMPU 检验．当 $\alpha>1/2$ 时，(3.79)是(3.78)的 UMP 检验．

（三）对比试验中的其他检验法

回到上一段中甲、乙两处理在 n 对试验单元中的比较问题．当试验误差服从正态分布时，相应的检验问题是(3.72)或(3.78)（其中 $a_0=0$），用一样本 t 检验去处理．若没有把握认为试验误差为正态，则必须用另外的方法去检验．下面介绍几个重要的方法．

1．*符号检验*　在有的情况下，同一对试验单元中甲、乙两处理的比较，不能用一个数量指标去表述．例如，让同一个人品尝两种酒，他的评价可能是“甲比乙好”，或“乙比甲好”，或“难于区别”以致不予回答．每对单元中的试验结果给出一个符号：

$+$：甲比乙好；$-$：乙比甲好；0：难区别，无结果．

根据各种符号的多少来检验“甲、乙一样”这个假设．

设 n 对单元的试验中，“$+$”号出现 n_+ 次，“$-$”号出现 n_- 次，其余为“0”．如果甲、乙一样，则在 $n'=n_++n_-$ 个非 0 结果中，每个有同等的机会是 $+$ 或 $-$．故在这个情况下，n_+ 的分布为二项分布 $B\left(n',\ \frac{1}{2}\right)$．若甲、乙两处理确有优劣之分，则每个结果取“$+$”号的概率 $p\neq\frac{1}{2}$．故若记 $X=n_+$，则所提问题转化为检验问题：X

$\sim B(n', p)$, $0\leqslant p\leqslant 1$, $H: p=\frac{1}{2}\leftrightarrow K: p\neq\frac{1}{2}$. 一个合适的检验为：

当$|X-n'/2|>C$时，否定"甲、乙一样"，不然就接受.

$$(3.80)$$

C之值要根据给定的水平α，用二项分布表去决定. 为使α是真实水平，必要时须用随机检验(即令$\varphi(X)=\gamma$当$|X-n'/2|=C$). 若n'很大，则可用大样本检验：利用当$p=1/2$而$n'\to\infty$时，$2(X-n'/2)/\sqrt{n'}$的分布收敛于$N(0, 1)$，可得出(3.80)中C的近似值为$\sqrt{n'}u_{\alpha/2}/2$. u_α由(3.54)确定.

有时，检验的目的是从"甲不优于乙"和"甲优于乙"中选择其一. 以前者为原假设，则问题成为：$X\sim B(n', p)$, $0\leqslant p\leqslant 1$, $H: p\leqslant 1/2\leftrightarrow K: p>\frac{1}{2}$. 这种问题在例3.1中就出现过，并在例3.9中证明它有UMP检验，其否定域为$\{X>C\}$这种形式.

这里讨论的检验通称**符号检验**，因为它是基于试验结果的符号. 从统计模型而言，它不过是二项分布参数检验的一个特例.

2. Fisher的置换检验　这种检验的思想和实施方法已在例3.3中仔细描述过了. 由于(3.5)中那2^{15}个值是按(3.4)的方式把符号作一切可能的置换而得到，故称为**置换检验**. 在此我们作几点补充说明.

一是例3.3所描述的模型，与导致假设(3.72)时所据的假定有些不同. 这里所指的还不是随机误差是否服从正态分布一点，而在于：在导致(3.72)的论述中，我们认为同一对试验单元的差别可忽略不计，而随机误差纯系由试验中其他种种不可控因素而来. 在例3.3中则相反，误差的来由正是因为同一对内的试验单元的条件有所差别，而其他种种不可控因素导致的误差则可忽略不计. 虽然如此，不难看出，在例3.3中的论证，即使在这里仍然有效. 因为，如甲、乙两处理无差别，而同一对内试验单元的分配又是等可能的，则甲、乙处在完全对称的地位. 比方说，在一次试验中发现$X_1=1.5$，则根据上述对称性，也有同等的可能$X_1=-1.5$. 因

此，例 3.3 那里的讨论仍可用于此处. 即使在更一般的场合，其中既有不可控的随机因素起作用而同一对内的试验单元也有所差别，例 3.3 所描述的置换检验程序同样可用.

另一点是在例 3.3 中我们曾指出，实施该例中所描述的检验法，其困难在于计算量太大. Fisher 自己及其他许多学者，都研究过这样的问题：当 n 很大时，可否找到一种近似的程序去实施置换检验，以大大简化计算？研究结果证明了：在很一般的条件下，这种简化程序不仅存在，且就是通常的 t 检验！这是一个很有意思的结果. 因为，一开始，t 检验是在一个很有局限性的模型(正态)中导出的. 通过这个途径，发现即使在远为广泛的模型下，只要试验次数足够大，t 检验仍是合用的. 因此可以说，置换检验的理论从一个侧面加强了 t 检验的地位. 无庸赘言，在 n 不大时，置换检验与 t 检验有显著的差别.

(四) 两样本 t 检验

设样本 $X_1, \cdots, X_m \sim N(a, \sigma^2)$, $Y_1, \cdots, Y_n \sim N(b, \sigma^2)$, 且全体样本独立. 这里 a、b 和 σ 都是未知参数，注意方差 σ^2 相同. 考虑检验问题

$$H: a=b \leftrightarrow K: a \neq b. \tag{3.81}$$

通常把 $N(a, \sigma^2)$ 和 $N(b, \sigma^2)$ 看作是两个总体的分布，而 X 样本和 Y 样本分别来自这两个总体，故(3.81)常称为两样本问题. 这种问题最常见的情况是两处理的比较，但不是成对比较：选择 $m+n$ 个条件比较均匀的试验单元. 随机地从其中抽取 m 个施加处理甲，其结果记为 $X_1, \cdots, X_m$，剩下的施加处理乙，其结果记为 $Y_1, \cdots, Y_n$. 若假定对每个处理而言，随机误差都服从均值为 0 的正态分布且有等方差，又设各试验单元的随机误差相互独立，则甲、乙两处理是否有差别的判定，就归结为检验问题(3.81).

现导出(3.81)的似然比检验. 记 $x=(x_1, \cdots, x_m)$, $y=(y_1, \cdots, y_n)$. 合样本 $(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$ 的似然函数为

$$f(x,\ y,\ a,\ b,\ \sigma)=(\sqrt{2\pi}\sigma)^{-(m+n)}$$
$$\times\exp\left[-\frac{1}{2\sigma^2}\left(\sum_1^m(x_i-\bar{x})^2+\sum_1^n(y_j-\bar{y})^2\right)\right]. \tag{3.82}$$

利用(3.74)和(3.75)式，易得

$$\begin{aligned}&\sup\{f(x,\ y,\ a,\ b,\ \sigma):\ -\infty<a,\ b<\infty,\ \sigma>0\}\\&=\sup\left\{(\sqrt{2\pi}\sigma)^{-(m+n)}\exp\left(-\frac{1}{2\sigma^2}S_1\right):\ \sigma>0\right\}\\&=[2\pi e/(m+n)]^{-(m+n)/2}S_1^{-(m+n)/2},\end{aligned}$$

其中 $S_1=\sum_1^m(x_i-\bar{x})^2+\sum_1^n(y_j-\bar{y})^2$. 同样，有

$$\begin{aligned}&\sup\{f(x,\ y,\ a,\ b,\ \sigma):\ -\infty<a=b<\infty,\ \sigma>0\}\\&=\sup\left\{(\sqrt{2\pi}\sigma)^{-(m+n)}\exp\left(-\frac{1}{2\sigma^2}S_2\right):\ \sigma>0\right\}\\&=[2\pi e/(m+n)]^{-(m+n)/2}S_2^{-(m+n)/2},\end{aligned}$$

此处 $S_2=\sum_1^m(x_i-z)^2+\sum_1^n(y_j-z)^2$，而 $z=(m\bar{x}+n\bar{y})/(m+n)$. 因此似然比为

$$LR(x,\ y)=(S_2/S_1)^{(m+n)/2}.$$

注意到 $m(\bar{x}-z)^2+n(\bar{y}-z)^2=mn(\bar{x}-\bar{y})^2/(m+n)$，以及 $S_2=S_1+m(\bar{x}-z)^2+n(\bar{y}-z)^2$，易见似然比是

$$T^*=|\bar{x}-\bar{y}|/\sqrt{S_1}$$

的严增函数. 由此知(3.81)的似然比检验有否定域$\{|T^*|>C\}$. 为确定 C，我们来证明：当原假设 $a=b$ 成立时，有

$$T_1=\sqrt{\frac{mn(m+n-2)}{m+n}}(\bar{X}-\bar{Y})/\sqrt{S_1}\sim t_{m+n-2}. \tag{3.83}$$

事实上，T 可写为

$$T_1=\left[\sqrt{\frac{mn}{m+n-2}}(\bar{X}-\bar{Y})/\sigma\right]\Big/\sqrt{\frac{1}{m+n-2}(S_1/\sigma^2)}. \tag{3.84}$$

由于两组样本独立，又由定理 1.1，$\bar{X}$与 $\sum_1^m(X_i-\bar{X})^2$ 独立，$\bar{Y}$ 与

$\sum_1^n(Y_j-\overline{Y})^2$独立. 故 $\overline{X}$, $\overline{Y}$, $\sum_1^m(X_i-\overline{X})^2$, $\sum_1^n(Y_j-\overline{Y})^2$ 四变量相互独立, 由此推出 $\overline{X}-\overline{Y}$ 与 S_1 独立. 又 $\sqrt{\frac{mn}{m+n-2}}(\overline{X}-\overline{Y})/\sigma$ $\sim N(0, 1)$(当 $a=b$ 时), $\sum_1^m(X_i-\overline{X})^2/\sigma^2\sim\chi_{m-1}^2$, $\sum_1^n(Y_j-\overline{Y})^2/(n-1)\sim\chi_{n-1}^2$ 且二者独立, 故 $S_1/\sigma^2\sim\chi_{m+n-2}^2$(定理1.1及系1.1). 于是由(3.84)及 t 分布的定义即得(3.83).

由(3.83), 得(3.81)的水平 α 的似然比检验有否定域

$$|T_1|>t_{m+n-2}(\alpha/2). \tag{3.85}$$

可以证明: 它是 UMPU 检验. (3.83)所定义的统计量 T_1 常称为**两样本 t 统计量**. 基于它的检验, 例如(3.85), 称为**两样本 t 检验**.

若检验问题是 $H: a-b=c \leftrightarrow K: a-b\neq c$($c$给定), 则只须在 T_1 的定义中, 把分子的 $\overline{X}-\overline{Y}$ 改为 $\overline{X}-\overline{Y}-c$, 其他无变化.

对检验问题

$$H: a-b\leqslant 0 \leftrightarrow K: a-b>0 \tag{3.86}$$

来说, 易算出当 $\bar{y}\geqslant\bar{x}$ 时, 有

$$\begin{aligned}&\sup\{f(x, y, a, b, \sigma): -\infty<a, b, <\infty, \sigma>0\}\\&=\sup\{f(x, y, a, b, \sigma: -\infty<a\leqslant b<\infty, \sigma>0\}\\&=[2\pi e/(m+n)]^{-(m+n)/2}S_1^{-(m+n)/2}.\end{aligned}$$

这时有 $LR(x, y)=1$. 若 $\bar{y}<\bar{x}$, 则首先注意, $\sup\{\sum_1^m(x_i-a)^2+\sum_1^n(y_j-b)^2: a\leqslant b\}$必在 $a=b$ 时达到. 事实上, 任取 $a<b$. 若 $b\leqslant\bar{x}$, 则

$$\sum_1^m(x_i-b)^2+\sum_1^n(y_j-b)^2>\sum_1^m(x_i-a)^2+\sum_1^n(y_j-b)^2.$$

若 $b>\bar{x}$, 则当 $a\leqslant\bar{y}$ 时, 改 a、b 都等于 $\bar{y}$; 当 $\bar{y}<a<\bar{x}$时改 b 为 a, 都可以使上述平方和增大. 由此得出上述结论, 从而知当 $\bar{y}<\bar{x}$ 时

$$\begin{aligned}&\sup\{f(x, y, a, b, \sigma): -\infty<a\leqslant b<\infty, \sigma>0\}\\&=\sup\{f(x, y, a, b, \sigma): -\infty<a=b<\infty, \sigma>0\}\end{aligned}$$

$$= [2\pi e/(m+n)]^{-(m+n)/2} S_2^{-(m+n)/2}.$$

总结上述 得

$$LR(x, y) = \begin{cases} 1, & \text{当 } \bar{y} \geqslant \bar{x}; \\ (S_2/S_1)^{(m+n)/2}, & \text{当 } \bar{y} < \bar{x}, \end{cases} \quad (3.87)$$

与例 3.12 与(3.75)相似，由此得出：若给定水平 $\alpha \leqslant \frac{1}{2}$，则(3.86)的似然比检验有否定域

$$T_1 > t_{m+n-2}(\alpha). \quad (3.88)$$

若 $\alpha > \frac{1}{2}$，则似然比检验将是随机化的. 因此在这个情况下我们不拘泥于似然比检验，而仍取以(3.88)为否定域的检验. 可以证明：这个检验是 UMPU 检验.

检验问题 $H: a-b \geqslant 0 \leftrightarrow K: a-b<0$ 不是新问题，因为可交换 a、b 的位置. 又若原假设为 $a-b \leqslant c$（c 给定），则只须用 $\bar{X}-\bar{Y}-c$ 代替(3.83)右边分子中的 $\bar{X}-\bar{Y}$，其余一切无变化.

前面我们假定了两总体的方差相同. 若方差不同且未知，则得到下面的问题：$X_1, \cdots, X_m \sim N(a, \sigma_1^2)$，$Y_1, \cdots, Y_n \sim N(b, \sigma_2^2)$，全部样本独立，$a$、$b$、$\sigma_1^2$、$\sigma_2^2$ 都未知，要检验假设(3.81)或(3.86). 这个问题世称 **Behrens-Fisher** 问题，是 Behrens 在 1929 年首先提出，Fisher 等在三十年代一系列文章中讨论过. 这个假设检验问题引起了很多著名学者的兴趣，有关的文献很多，提出了种种检验方法，小样本大样本的都有. 我们将在 § 4.2 中结合讨论 Fisher 的信任推断法，去讨论 Fisher 提出的一种检验法.

（五）两样本问题其他检验法

当随机误差不服从正态分布时，两处理的比较问题就需要更广的提法，并使用相应的检验方法. 这方面的理论和方法很多，但大都很专门. 这里只能很简略地讨论一下.

1. 置换检验法　这个方法的思想及具体作法与例 3.3 相似. 把两处理的全部试验结果，即 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$，改记为

$Z_1, \cdots, Z_{m+n}$. 如果两处理甲、乙无差别，则 $Z_1, \cdots, Z_{m+n}$ 之间的差别不是由于处理的不同而来，而是由于这 $m+n$ 个试验单元的分配方法而来. 在 $m+n$ 个试验单元中分配 m 个给处理甲，不同的方法有 $\binom{m+n}{m}$ 种. 若在每种分配方法之下都计算以下的量：

$$g=\frac{1}{m}(\text{甲处理试验值之和})-\frac{1}{n}(\text{乙处理试验值之和}).$$

它等于 $Z_1, \cdots, Z_{m+n}$ 中的 m 个的平均减去剩下的 n 个的平均. 那 m 个则要看试验单元是如何分配的. 这样，一共能得到 $N=\binom{m+n}{m}$ 个值：

$$g_1,\ g_2,\ \cdots,\ g_N,\ |g_1|\geqslant|g_2|\geqslant\cdots\geqslant|g_N|. \tag{3.89}$$

这与(3.5)那 2^{15} 个值的作用相似. 在"甲、乙两处理效果一样"的原假设成立时，(3.89)中那 N 个值中的每一个，有同等出现的机会 $1/N$. 就实际样本算出 g 之值，即 $g^*=\overline{X}-\overline{Y}$. 如果原假设不成立，$|g^*|$ 倾向于取较大之值. 因此，给定水平 α 后，用下面的方法检验：

$$\text{当}\ |g^*|>|g_{[N\alpha]}|\ \text{时否定原假设，不然就接受.} \tag{3.90}$$

如果原假设是"处理甲不优于乙"，对立假设是"处理甲优于乙"，则只有在 g^*(而不是 $|g^*|$)大时才导致否定原假设. 这时，(3.89)中那 N 个值要排列为 $h_1\geqslant h_2\geqslant\cdots\geqslant h_N$，只有在 $g^*>h_{[N\alpha]}$ 时才否定原假设. 这个检验的"置换"特点更明显：它是基于 $m+n$ 个原始试验值作一切可能的置换而产生的. 在一样本情况下对置换检验所作的两点说明，仍完全适用于此处. 特别，当 m、n 都很大时，上述置换检验接近于两样本 t 检验. 有如在一样本的情况，这个性质从一个侧面加强了两样本 t 检验的地位.

2. Wilcoxon 秩和检验　我们再看看假设(3.81). 若记 $\theta=b-a$，而 $F(x)$ 为第一总体的分布 $N(a, \sigma^2)$，则第二总体的分布 $N(b, \sigma^2)$ 可表为 $F(x-\theta)$. 如果我们免除掉 $F(x)$ 为正态分布函数这一假定，而把它当作完全未知的，则得到这样一个问题：X_1,

…, $X_m \sim F(x)$, $Y_1, \cdots, Y_n \sim F(x-\theta)$，要检验假设

$$H: \theta=0 \leftrightarrow K: \theta\neq 0 (F \text{ 完全未知}). \tag{3.91}$$

这是一个典型的非参数检验问题，因为样本$(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$的分布族为$\{F(x_1)\cdots F(x_m)F(y_1-\theta)\cdots F(y_n-\theta): F$ 任意$\}$，这个分布族不能用有限个实参数去刻划．如果把 $X_1, \cdots, X_m$ 和 $Y_1, \cdots, Y_n$ 分别看作是甲、乙两个处理的试验值，则(3.91)就是两处理效果无差别的检验问题，但此处对随机误差的分布已不再假定为正态的．

暂设 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$，这 $m+n$ 个值两两不同．把它们按大小排列，结果为

$$Z_1<Z_2<\cdots<Z_N, \ N=m+n. \tag{3.92}$$

每个 Y_i 必是序列(3.92)中的某一个．若 $Y_i=Z_{R_i}$，则称 Y_i 在合样本 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ 中的秩为 R_i．记

$$W=R_1+\cdots+R_n, \tag{3.93}$$

它称为 **Wilcoxon 的两样本秩和统计量**，是 Wilcoxon 在 1945 年一项工作中引进的．

现在这样推理：每个 R_i 都可取 $1, 2, \cdots, N$ 为值，这 N 个数的平均为 $\frac{1}{N}(1+2+\cdots+N)=\frac{N+1}{2}$．若 $\theta=0$，则全部样本来自同一分布，每个都不占特殊地位．故 W 所取之值应在其平均值 $n\cdot\frac{N+1}{2}$ 附近．反之，若 $\theta>0$，则 Y 样本倾向于比 X 样本大，而 R_i 倾向于取 $1, 2, \cdots, N$ 中较大一端的值．因此，W 取值将倾向于比 $n(N+1)/2$ 大．反之，若 $\theta<0$，则 W 倾向于取比 $n(N+1)/2$ 为小的值．总之，当 $\theta\neq 0$ 时，$|W-n(N+1)/2|$ 倾向于取大值．这种分析导致如下的检验法：

$$\text{当} |W-n(N+1)/2|>C \text{ 时否定 } \theta=0, \text{不然就接受}. \tag{3.94}$$

C 的确定在原则上可以解决：当原假设 $\theta=0$ 成立时，合样本独立同分布．由此根据对称性的考虑，易知$(R_1, \cdots, R_n)$的分布为

$$P(R_1=r_1,\ \cdots,\ R_n=r_n)$$

$$=\begin{cases}(N(N-1)\cdots(N-n+1))^{-1}, & \text{当 } r_1,\ \cdots,\ r_n \text{ 为彼此不同的, 不超过 } N \text{ 的自然数;}\\ 0, & \text{其他.}\end{cases} \tag{3.95}$$

由此就不难形式地写出 W 的分布，从而根据给定的水平 α 去确定 (3.94) 中的 C(必要时随机化，或略修改 α 之值). 对较小的 m, n 这是可行的且已造了表. 但当 m、n 较大时计算很复杂，因此得依靠极限定理. 可以证明：在原假设 $\theta=0$ 成立之下，当 m、n 都 $\to\infty$ 时，统计量

$$(W-n(N+1)/2)\Big/\sqrt{\frac{1}{12}mn(N+1)} \tag{3.96}$$

的分布收敛于标准正态分布 $N(0, 1)$. 由此，当 m、n 都较大时，对给定的水平 α，可以按公式

$$C\approx u_{\alpha/2}\sqrt{\frac{1}{12}mn(N+1)}, \tag{3.97}$$

近似决定(3.94)中的 C.

如果假设是单边的，即：

$$H:\theta\leqslant 0\leftrightarrow K:\theta\nleqslant 0(F \text{ 任意}), \tag{3.98}$$

则将双边检验(3.94)修改为单边的，即：

$$\text{当 } W-n(N+1)/2>C \text{ 时否定 } H, \text{不然就接受.} \tag{3.99}$$

给定 α 后，C 的值也可通过 W 的分布(在 H 成立时)定出. 在 m、n 较大时，则利用统计量(3.96)的极限分布. 它给出(3.99)中 C 的近似值

$$C\approx u_{\alpha}\sqrt{\frac{1}{12}mn(N+1)}. \tag{3.100}$$

检验(3.94)和(3.99)都称为 Wilcoxon 的两样本秩和检验. 它是 Wilcoxon 在 1945 年提出的一种重要的秩检验. 一般，秩检验是指这样一类检验，在其中只利用样本的秩. 秩检验是最重要的一类非参数检验，有深刻的理论和广泛的应用.

乍一看可能会觉得，这种秩检验效率不会高，因为它只利用了样本中的大小关系而完全忽略了其具体数值，其实不然．近代关于秩检验的大样本理论证明了，一般秩检验至少在样本大小较大时，与传统的参数检验并不逊色．拿 Wilcoxon 检验与两样本 t 检验的比较来说，即使当随机误差分布 F 确为正态时，Wilcoxon 检验的效率相对于 t 检验也达到 $3/\pi\approx0.95$（样本大小较大时）．对别的 F，这个相对效率可任意大，且总不会低于 0.864.

以上我们假定了合样本 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ 彼此不同，因而 Y_i 的秩 R_i 可唯一定出．若假定 F 处处连续，这一事实将以概率 1 成立，这时不存在问题．当 F 不连续时，合样本中可能出现相同的，即所谓"结"的问题．这是秩理论中一个很细致的问题，在此不仔细讨论了．

最后我们指出：在文献中 Wilcoxon 检验也常被称为 Mann-Whitney 检验，或 Wilcoxon-Mann-Whitney 检验．因为后两位作者在 1947 年提出了另一个统计量，它与 W 比只相差一个常数，二者所导出的检验当然完全相同．

（六）正态分布方差的检验

1．一样本情况　设 $X_1, \cdots, X_n\sim N(a, \sigma^2)$，$a$ 和 σ 都是未知参数．先考虑检验问题

$$H: \sigma^2=\sigma_0^2 \leftrightarrow K: \sigma^2\neq\sigma_0^2. \tag{3.101}$$

这里 $\sigma_0>0$ 为给定的已知值.

下面先定出(3.101)的水平 α 的似然比检验，似然函数为(3.73)．利用(3.74)和(3.75)，不难算出似然比为

$$LR(x_1, \cdots, x_n)=(e\xi)^{-n/2}e^{\xi/2},\quad \xi=\frac{1}{\sigma_0^2}\sum_1^n(x_i-\bar{x})^2. \tag{3.102}$$

注意到函数 $g(x)=x^{n/2}e^{-x/2}$ 在 $x>0$ 处先降后升，且由定理 1.1 知，当 $\sigma^2=\sigma_0^2$ 成立时，ξ 的分布为自由度 $n-1$ 的 χ^2 分布，即 χ_{n-1}^2，不难得到(3.101)的水平 α 的似然比检验为：

$$\text{当 } k_1 \leqslant \xi \leqslant k_2 \text{ 时接受原假设,不然就否定。} \tag{3.103}$$

其中 k_1, k_2 为下述方程组的解:

$$\begin{cases} k_1^{n/2} e^{-k_1/2} = k_2^{n/2} e^{-k_2/2}; \\ P(\chi_{n-1}^2 < k_1) + P(\chi_{n-1}^2 > k_2) = \alpha. \end{cases} \tag{3.104}$$

这个似然比检验不是 UMPU 的检验, 但与之相去不远. 可以证明: 若把(3.104)的第一方程中的 n 改为 $n-1$ 而解出 k_1、k_2, 则(3.103)将是 UMPU 检验.

方程(3.104)的解不易得到. 一般就取

$$k_1 = \chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right), \quad k_2 = \chi_{n-1}^2\left(\frac{\alpha}{2}\right), \tag{3.105}$$

而采用检验(3.103). 这样得到的检验仍有水平 α, 与 UMPU 检验相去不远. 而 k_1、k_2 可由 χ^2 分布表上查得.

单边假设 $\sigma^2 \leqslant \sigma_0^2$ 已在(3.55)那里提到过, 在该处指出了其 UMP 检验.

2. 两样本情况 设 $X_1, \cdots, X_m \sim N(a, \sigma_1^2)$, $Y_1, \cdots, Y_n \sim N(b, \sigma_2^2)$, a、b、σ_1^2 和 σ_2^2 都是未知参数, 又合样本 $X_1, \cdots, X_m$, $Y_1, \cdots, Y_n$ 相互独立. 考虑检验问题

$$H: \sigma_1^2 \leqslant \sigma_2^2 \leftrightarrow K: \sigma_1^2 > \sigma_2^2. \tag{3.106}$$

另一个假设检验问题 $\sigma_1^2 \geqslant \sigma_2^2 \leftrightarrow \sigma_1^2 < \sigma_2^2$ 可通过更换 X, Y 的地位转化为(3.106). 另外,还有双边检验问题

$$H: \sigma_1^2 = \sigma_2^2 \leftrightarrow K: \sigma_1^2 \neq \sigma_2^2. \tag{3.107}$$

(3.106)的似然比检验容易推得,此处只写出其结果: 记

$$F = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2 \Big/ \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \bar{Y})^2. \tag{3.108}$$

在原假设成立时, F 服从自由度为 $m-1$ 和 $n-1$ 的 F 分布, 即 $F_{m-1,n-1}$. 定义 $F_{m-1,n-1}(\alpha)$, 使

$$P(F_{m-1,n-1} > F_{m-1,n-1}(\alpha)) = \alpha. \tag{3.109}$$

$F_{m-1,n-1}(\alpha)$的值可在 F 分布表上查到. (3.106)的水平 α 的似然比检验为:

$$当\ F>F_{m-1,n-1}(\alpha)时否定原假设，不然就接受 \qquad (3.110)$$

(与以前几个情况相似，当 $\alpha<1/2$ 时，(3.110)确为似然比检验，而当 $\alpha>1/2$ 时则略加调整)可以证明，它是(3.106)的水平 α 的 UMPU 检验。

(3.107)的似然比检验要复杂些。利用(3.74)和(3.75)易算出其似然比为

$$LR(x_1, \cdots, x_m, y_1, \cdots, y_n)=C(1+F^*)^{n/2}\left(1+\frac{1}{F^*}\right)^{m/2}. \qquad (3.111)$$

此处 $F^*=\frac{m-1}{n-1}F$，而 F 由(3.108)给出，C 为一个只与 m, n 有关的常数。不难验证(证明留给读者)：(3.111)右边作为 F^* 的函数，在 $0<F^*<\infty$ 内先降后升。由此得出(3.107)的水平 α 的似然比检验为：

$$当\ f_1\leqslant F\leqslant f_2\ 时接受原假设，不然就否定。\qquad (3.112)$$

其中 f_1 和 f_2 是下述方程组的解：

$$\begin{cases}\left(1+\dfrac{m-1}{n-1}f_1\right)^{n/2}\left(1+\dfrac{n-1}{m-1}\dfrac{1}{f_1}\right)^{m/2} \\ \quad =\left(1+\dfrac{m-1}{n-1}f_2\right)^{n/2}\left(1+\dfrac{n-1}{m-1}\dfrac{1}{f_2}\right)^{m/2}; \\ P(F_{m-1,n-1}<f_1)+P(F_{m-1,n-1}>f_2)=\alpha. \end{cases} \qquad (3.113)$$

这似然比检验不是 UMPU 检验，但与之相去不远。可以证明：若把(3.113)第一方程中的指数 $m/2$ 和 $n/2$ 分别改为 $(m-1)/2$ 和 $(n-1)/2$ 而解出 f_1 和 f_2，则(3.112)将是 UMPU 检验。

方程(3.113)不易解。一般就取

$$f_1=F_{m-1,n-1}\left(1-\frac{\alpha}{2}\right),\ f_2=F_{m-1,n-1}\left(\frac{\alpha}{2}\right), \qquad (3.114)$$

而采用检验(3.112)。这可由 F 分布表上查出。这样定出的检验有水平 α，且接近 UMPU 检验。

§3.7 序贯概率比检验

(一)序贯检验

到目前为止,我们在对一个假设进行检验时,所用样本的样本大小是一个固定的常数,与样本本身的取值无关. 凡是基于这种样本的检验方法,就称为**固定样本检验**. 这个固定的样本大小 n 的选定,可以是基于某种考虑.比方说,为了使检验有足够的功效,n 至少需要多大. 也有可能,样本就是一批现成的资料. 这时 n 是一个"既成事实"而并非出自有意的安排. 但这些都无关宏旨. 如在例 3.1 中,为要确定一批产品是否该接受,先决定一个 n,然后从这批产品中抽出 n 个,以进行检验.

与此对立的一种做法是:我们不是一劳永逸地决定一个样本大小 n,而是在抽样的过程中去决定它. 确切地说,我们先定下第一批抽样多少个,例如 n_1 个. 抽样得到样本 $X_1, \cdots, X_{n_1}$ 后,我们根据这些样本的具体值去决定:或者是抽样到此为止而作出一个决定(接受或否定原假设),或者我们认为,由已得的样本 $X_1, \cdots, X_{n_1}$ 尚难于作出这种决定,而需要再抽一些样,抽的个数 n_2 也取决于 $X_1, \cdots, X_{n_1}$ 的值. 一般地,在每一阶段的抽样完成后,我们都面临上述抉择:或者停止抽样,或继续抽,抽多少也由到当时为止已有的样本去决定.凡是按这样的方式去进行抽样以作检验的,就叫做**序贯**检验. 形式上,固定样本检验可视为序贯检验的一个特例.

使用序贯检验的原因,一般不外乎以下两种:

1. 在有的情况下,不论你事先指定的样本大小多大,所抽得的样本使我们感到难于作出决定. 如检验正态分布 $N(\theta, 1)$中关于 θ 的假设 $\theta \leqslant 0$. 如果由样本算出的 $\overline{X}$ 就等于 0,或 $|\overline{X}|$ 极小,我们会觉得,不论说 $\theta \leqslant 0$ 或 $\theta > 0$ 都嫌证据不足. 若勉强下一结论就很可能出错. 又如在例 3.1 中,若样本中所含废品数很多或很少,我们感到较有把握作出决定. 反之,若废品数不多不少,则

感到下结论把握不大．这时可能有必要再抽出些产品作检验．

2．另一个原因是为节省试验次数，因而节省费用．一般地这可以理解为：为了使所作的统计推断达到一定的性质(例如，为使一个水平 α 的检验的功效至少等于 β)，使用序贯抽样平均说来，可能比使用固定样本的抽样次数少．一个简单但不甚典型的例子如下：设在例 3.1 中，买卖双方决定抽 20 件产品，若其中废品数不超过 2 就接收该批产品，否则就拒收．按固定样本作法须抽 20 件产品去检验．但我们也可以一个一个地抽，直到抽出 3 个废品，或抽满 20 件为止．这时，在某些情况下抽样的次数就可以小于 20．

这种序贯抽样方法显然不止适用于假设检验，对参数估计和其他统计推断问题也适用．其实，这种作法在种种领域以至日常生活中也很常见．例如，在掌握情报不够时，推迟对一件事情采取行动，以待有更充分的情报时再说．序贯方法的主要缺点也正在于其“序贯”性．它使应用者在事先对“到底要试验到何时”心里没有底，而产生心理上的压力．另外，在有些情况下，分阶段抽样在组织和实施上也带来不少麻烦．以此之故，总的说来，在目前序贯方法的应用还不多．

序贯检验最早的一个重要例子是产品验收中的所谓**复式抽样方案**．一批产品的废品率 p 未知．买卖双方经过协商定下 4 个数：α、β、p_α、p_β，$0<\alpha$，$\beta<1$，$0<p_\alpha<p_\beta<1$．要制定一种抽样验收方案，使：1．当废品率 $p\leqslant p_\alpha$ 时，这批产品被拒收的概率不超过 α．2．当废品率 $p\geqslant p_\beta$ 时，这批产品被接受的概率不超过 β．Duguè 和 Romig 设计的复式抽样方案如下：确定五个数 n_1，n_2，c_1，c_2，c．先抽 n_1 个产品去检验．若其中的废品个数 $X_1\leqslant c_1$，则接受该批产品，若 $X_1\geqslant c_2$ 则拒收．若 $c_1<X_1<c_2$，则再抽 n_2 个产品去检验．以 X 记所抽出的 n_1+n_2 个产品中的废品数．若 $X\leqslant c$，则接受该批产品，否则就拒收．满足上述 1.2. 两个条件的固定(单式)抽样方案(即：定下两个数 n，c．抽 n 个产品．视其中废品数是否 $\leqslant c$ 而决定是否接受该批产品)也可以找到，但所需抽样次数一般比复式方案多．

上面介绍的复式抽样方案是一种**两阶段抽样**，且第二阶段的抽样次数 n_2 也定了(一般的两阶段抽样，第二阶段抽样次数与第一阶段所得样本有关). 一般，对任何自然数 m，可有 **m 阶段抽样**. 每阶段抽样个数可以多于 1，且可与以前阶段所得样本值有关. 若对 m 不加限制，就是所谓**纯序贯抽样**. 这时，不论指定的 m 多大，有可能(视所得样本而定)在经过 m 阶段的抽样后，还须继续再抽. 这种纯序贯抽样的一个重要例子是著名统计学家 A. Wald 在二次大战期间发展起来的，叫序贯概率比检验. Wald 是为适应战时大量验收军需品的要求而提出他的方法的. 战后 1947 年，他发表了«Sequential Analysis»这本著作，介绍了他的方法.

前已提到，序贯方法目前应用尚不多. 在这不多的应用中，Wald 的序贯概率比检验占有突出的地位，以此之故，我们在下面几段中简要地介绍一下这种检验，但不过多地深入其细节.

(二)序贯概率比检验的定义

设总体分布有概率函数 $f(x, \theta)$，参数 θ 只有两个可能值: θ_1 和 θ_2 要检验假设

$$H:\theta=\theta_1 \leftrightarrow K:\theta=\theta_2. \tag{3.115}$$

若给定了样本大小 n，且抽样得到独立随机样本 $x_1, \cdots, x_n$. 则根据 NP 基本引理，(3.115)的 UMP 检验有否定域$\{\prod_{i=1}^{n}[f(x_i, \theta_2)/f(x_i, \theta_1)]>C\}$，$C$ 取决于给定的水平 α (暂不考虑随机检验).

把 C 这个值作为决定是否接受 H 的截然界线，使人觉得太绝对化. 为了克服这一点，Wald 引进了一种序贯检验. 下面给出其形式定义.

定义 3.8 在上述记号下，确定一种序贯检验程序如下: 指定常数 A, B, $A<B$, 样本 $x_1, x_2, \cdots$一个一个地抽(每次一个). 若在得到 $x_1, \cdots, x_{n-1}$ 后抽样尚不能停止，则再抽 x_n. 然后计算 $l_n=\prod_{i=1}^{n}[f(x_i, \theta_2)/f(x_i, \theta_1)]$，且当 $l_n\leqslant A$ 时接受原假设 H，当 $l_n\geqslant B$ 时

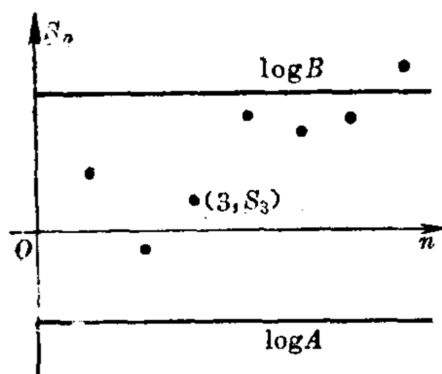
否定 H. 若 $A<l_n<B$, 则继续抽样得 x_{n+1}. 算出 l_{n+1} 再按以上规则处理. 这个检验称为**序贯概率比检验**, 简称 SPRT(是 Sequential Probability Ratio Test 的缩写).

定义中的常数 A、B, 要由一定的条件去确定, 见(三). SPRT 的直观意义很清楚: 只有当概率比 l_n 的值比较显著地偏于一边时才作决定, 不然就继续抽样, 直到这种情况出现时为止.

引进记号

$$Z_i=\log[f(x_i,\ \theta_2)/f(x_i,\ \theta_1)],\ S_n=\sum_{i=1}^{n} Z_i. \tag{3.116}$$

则 SPRT 可表为: 当 $S_n\leqslant\log A$ 时接受 H, 当 $S_n\geqslant\log B$ 时否定 H, 而当



$$\log A<S_n<\log B \tag{3.117}$$

时继续观察 x_{n+1}. 如在 $(n,\ S_n)$ 的直角坐标系中作两条与 n 轴平行的直线如图所示, 把平面 ($n\geqslant0$ 的部分) 割成三部分. 在此坐标系中标出点 $(n,\ S_n)$ 如图. 则只要点继续落在中间的带形域内, 就要继续抽样, 什么时候从上端(下端)越出这个区域, 就否定(接受)原假设 H.

举两个简单例子.

例 3.14 设总体分布为两点分布

$$P_p(X=1)=p,\ P_p(X=0)=1-p. \tag{3.118}$$

这相当于从一批有废品率 θ 的产品中作有放回(若批中产品数很大, 也可以是无放回的)的抽样, 并规定抽得废品时 $X=1$, 否则为 0. 设 $0<p_1<p_2<1$. 要求检验问题

$$H: p=p_1 \leftrightarrow K: p=p_2 \tag{3.119}$$

的 SPRT. 此处有 $f(x,\ p)=p^x(1-p)^{1-x}$, 故 $Z_i=X_i\log\dfrac{p_2}{p_1}+(1-X_i)\log\dfrac{1-p_2}{1-p_1}=X_i\log\dfrac{p_2(1-p_1)}{p_1(1-p_2)}+\log\dfrac{1-p_2}{1-p_1}$. 因此, 若记

$$A_n=\left(A-n\log\frac{1-p_2}{1-p_1}\right)\bigg/\log\frac{p_2(1-p_1)}{p_1(1-p_2)},$$

$$B_n=\left(B-n\log\frac{1-p_2}{1-p_1}\right)\bigg/\log\frac{p_2(1-p_1)}{p_1(1-p_2)},$$

则 SPRT 为：当 $\sum\limits_{i=1}^{n}X_i\leqslant A_n$ 时接受 H，$\sum\limits_{i=1}^{n}X_i\geqslant B_n$ 时否定 H，当 $A_n<\sum\limits_{i=1}^{n}X_i<B_n$ 时继续抽样.

例 3.15 总体分布为 $N(\theta,\ 1)$. 这时

$$f(x,\ \theta)=\frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2},$$

有 $Z_i=\frac{1}{2}(X_i-\theta_1)^2-\frac{1}{2}(X_i-\theta_2)^2=(\theta_2-\theta_1)X_i-\frac{1}{2}(\theta_2^2-\theta_1^2)$.
注意到 $\theta_2-\theta_1$，知检验问题(3.115)的 SPRT 与例 3.14 的 SPRT 有相同形式，但 A_n 和 B_n 分别改为：

$$A_n=\left(A+\frac{n}{2}(\theta_2^2-\theta_1^2)\right)/(\theta_2-\theta_1),$$

$$B_n=\left(B+\frac{n}{2}(\theta_2^2-\theta_1^2)\right)/(\theta_2-\theta_1).\qquad(3.120)$$

因为 SPRT 是从修改由 NP 基本引理所决定的检验得[illegible]因此有理由期望，这个检验具有某种优良性. 这一点确[illegible]实. 为了叙述这个结果，我们用 N 来表示当采用 SPRT 时[illegible]须的抽样次数. 注意 N 不是一个固定常数而是一个随机变[illegible]因为，抽样何时终止，要取决于抽得的样本，而样本是随机的. 当参数 θ 取值 θ_1 或 θ_2 时，平均抽样次数为 $E_{\theta_1}(N)$ 和 $E_{\theta_2}(N)$. 除此以外，还要考虑两类错误的概率. 以 SPRT$(A,\ B)$ 记问题(3.115)的、由定义 3.8 所确定的序贯概率比检验，它犯第一、二类错误的概率，分别记为 α 和 β. 设 $0<\alpha<1,\ 0<\beta<1$. 有下面的定理：

定理 3.6 设 φ 是问题(3.115)的任一个检验(序贯的或非序贯的)，其抽样次数记为 N^*，而其犯第一、二类错误的概率分别记为 α^*，β^*. 如果 $\alpha^*\leqslant\alpha$，$\beta^*\leqslant\beta$，则必有 $E_{\theta_1}(N^*)\geqslant E_{\theta_1}(N)$，$E_{\theta_2}(N^*)\geqslant E_{\theta_2}(N)$.

这个定理的意思是：在一切其犯两类错误的概率分别不超过α和β的检验类中，以SPRT的平均抽样次数最少．这个定理首先由Wald猜测其成立，并由他和Wolfowitz予以证明．证明细节不能在此给出了．

（三）边界值 A，B 的确定

问题是这样：给了α、β，$0<\alpha$，$\beta<1$．要决定A、B之值，使SPRT(A, B)犯一、二类错误的概率分别等于α和β．

这问题的确切解答只对个别分布族作出来了，且形式也很复杂．Wald在其前引著作中，提出了一个决定A、B的近似公式．下面来推导这个公式．为确定计，设$f(x, \theta)$是概率密度．记

$$T_n=\{(x_1, \cdots, x_n): A<\prod_{i=1}^{r}[f(x_i, \theta_2)/f(x_i, \theta_1)]<B,$$

$$r=1, \cdots, n-1; \prod_{i=1}^{n}[f(x_i, \theta_2)/f(x_i, \theta_1)]\geqslant B\}.$$

$$U_n=\{(x_1, \cdots, x_n): A<\prod_{i=1}^{r}[f(x_i, \theta_2)/f(x_i, \theta_1)]<B,$$

$$r=1, \cdots, n-1; \prod_{i=1}^{n}[f(x_i, \theta_2)/f(x_i, \theta_1)]\leqslant A\}.$$

T_n是这样一些$(x_1, \cdots, x_n)$的集：前$n-1$次抽样为止一直作不出决定，而在第n次抽样后决定否定$H:\theta=\theta_1$．因此

$$\alpha=\sum_{n=1}^{\infty}\int_{T_n}\left(\prod_{i=1}^{n}f(x_i, \theta_1)\right)dx_1\cdots dx_n. \tag{3.121}$$

因为在T_n上有$\prod_{i=1}^{n}f(x_i, \theta_1)\leqslant\frac{1}{B}\prod_{i=1}^{n}f(x_i, \theta_2)$，有

$$\alpha\leqslant\frac{1}{B}\sum_{n=1}^{\infty}\int_{T_n}\left(\prod_{i=1}^{n}f(x_i, \theta_2)\right)dx_1\cdots dx_n=(1-\beta)/B. \tag{3.122}$$

下面的直观想法给人这样的希望：(3.122)中的$\leqslant$号两边的量差别不大．这是因为，由$\prod_{i=1}^{n-1}[f(x_i, \theta_2)/f(x_i, \theta_1)]$到$\prod_{i=1}^{n}[f(x_i, \theta_2)/f(x_i, \theta_1)]$，多了一个因子．根据$T_n$的定义，前一乘积还小于$B$，

而后一乘积就大于 B 小，因此它不大会比 B 大很多。类似地有

$$1-\alpha=\sum_{n=1}^{\infty}\int_{U_n}\left(\prod_{i=1}^{n}f(x_i,\ \theta_1)\right)dx_1\cdots dx_n$$

$$\geqslant\frac{1}{A}\sum_{n=1}^{\infty}\int_{U_n}\left(\prod_{i=1}^{n}f(x_i,\ \theta_2)dx_1\cdots dx_n=\beta/A.\right. \tag{3.123}$$

由(3.122)，(3.123)，得出

$$\beta/(1-\alpha)\leqslant A<B\leqslant(1-\beta)/\alpha. \tag{3.124}$$

此式给出了 A 的下界和 B 的上界。根据前面所提的直观理由，可以希望(3.122)和(3.123)"大致上"是等式。由此得出 A、B 的近似值：

$$A\approx\beta/(1-\alpha),\ B\approx(1-\beta)/\alpha. \tag{3.125}$$

暂记 $A_1=\beta/(1-\alpha)$，$B_1=(1-\beta)/\alpha$。把 SPRT$(A_1,\ B_1)$ 作为近似解。我们要问：这个近似解，即SPRT$(A_1,\ B_1)$，犯一、二两类错误的概率 α_1 和 β_1 与原来给定的值 α 和 β 比较如何。按(3.124)(注意(3.124)不含近似成分)，应有

$$\beta_1/(1-\alpha_1)\leqslant A_1=\beta/(1-\alpha),\ (1-\beta_1)/\alpha_1\geqslant B_1=(1-\beta)/\alpha. \tag{3.126}$$

由(3.126)不难推得

$$\alpha_1\leqslant\alpha/(1-\beta),\ \beta_1\leqslant\beta/(1-\alpha),\ \alpha_1+\beta_1\leqslant\alpha+\beta. \tag{3.127}$$

因为 α，β 一般接近于 0，(3.127)的前两式表明，采用近似检验 SPRT$(A_1,\ B_1)$，其犯两类错误概率 α_1 和 β_1，即使比预定值 α，β 有所增加，但增加量也很小。(3.127)最后一式并指出：两类错误概率之和只能下降而不会增加。因此，使用近似值(3.125)可能的重要后果是：由于使错误概率不必要地过于缩小而增大平均抽样次数。例如，在问题(3.119)中若取 $p_1=0.05$，$p_2=0.17$，$\alpha=0.05$，$\beta=0.10$，则由近似公式(1.125)所确定的 SPRT$(A_1,\ B_1)$，其犯两类错误的概率分别为 $\alpha_1=0.031$ 和 $\beta_1=0.099$。后一值与预定值 0.10 很接近，但 α_1 与 $\alpha=0.05$相去则较远。

(四)复合假设的情况

在实际应用中，像(3.115)那样，原假设和对立假设都只包含一点的情况不多见，一般是参数 θ 可以在一个区间里取值. 在这种情况下，SPRT 如何用？本段来讨论一下这个问题.

设总体分布有概率函数 $f(x, \theta)$，θ 在一区间内取值，而 θ_0 是其一内点. 考虑检验问题

$$H:\theta\leqslant\theta_0 \leftrightarrow K:\theta>\theta_0. \tag{3.128}$$

通常，当参数值 θ 落在 θ_0 附近(离 θ_0 很近)时，不论是接受还是否定 H，从实际角度看都没有多大差别. 比方说 $\theta_0=0$. 当 $\theta=-10^{-10}$ 时，否定 H 也不一定很错；当 $\theta=10^{-10}$ 时，接受 H 也不一定很错. 在这种情况下往往可以指定两个数 θ_1、θ_2，$\theta_1<\theta_0<\theta_2$，使当 $\theta\leqslant\theta_1$ 时，否定 H 是严重错误；当 $\theta\geqslant\theta_2$ 时，接受 H 是严重错误. 至于当 $\theta_1<\theta<\theta_2$ 时，无论是接受或否定 H，情况都没有什么重大影响. 这样一来，原来的检验问题(3.128)可以用更切实际的检验问题

$$H:\theta\leqslant\theta_1 \leftrightarrow K:\theta\geqslant\theta_2 \tag{3.129}$$

去代替之.

再进一步看：在 $\theta\leqslant\theta_1$ 的范围内，唯有 θ_1 这个点与对立假设 $\theta\geqslant\theta_2$ 最接近. 因此它在某种程度上可取作原假设的代表. 类似地，θ_2 可取作对立假设的代表. 这样，我们可以从考虑简单的假设检验问题(3.115)入手，去处理(1.129)的问题. 更确切地说，我们对(3.115)找出一个较好的检验，期望它即使作为(3.129)的检验仍是较好的. 既然按定理 3.6，SPRT 是(3.115)的一个优良检验，那么，有理由期望，它也是(3.129)的一个优良检验. 这就实现了把 SPRT 用于复合假设检验的想法.

例如，对产品验收的例 3.14，若买卖双方商定的废品率界限为 0.05，则理应检验 $p\leqslant0.05 \leftrightarrow p>0.05$. 但双方可能认为，在 $0.045<p<0.055$ 的范围内，怎么处理都不算大错. 于是可以用 $p=0.045 \leftrightarrow p=0.055$ 来代替原先的检验问题.

但在这样做时，有一个重要问题要解决. 设用问题(3.115)代替(3.129)而使用 SPRT(A, B)，且 A, B 的选择要使 SPRT(A, B)作为(3.115)的检验时，其犯两类错误的概率分别不超过 α 和 β，则当把 SPRT(A, B)作为(3.129)的检验时，这个性质能否维持？更确切地说，若以 $\beta(\theta)$ 作为 SPRT(A, B)的功效函数，即

$$\beta(\theta)=\sum_{n=1}^{\infty}\int_{T_n}\left(\prod_{i=1}^{n}f(x_i,\ \theta)\right)dx_1\cdots dx_n, \tag{3.130}$$

则是否能成立

$$\beta(\theta)\leqslant\alpha \text{ 当 } \theta\leqslant\theta_1,\ \beta(\theta)\geqslant 1-\beta \text{ 当 } \theta\geqslant\theta_2. \tag{3.131}$$

注意，按 A, B 的取法，我们有

$$\beta(\theta_1)=\alpha,\ \beta(\theta_2)=1-\beta. \tag{3.132}$$

由(3.132)可知，如果 $\beta(\theta)$ 为 θ 的非降函数，则(3.131)必成立. 因此，关于是否能用(3.115)的 SPRT 去检验(3.129)的问题，就归结为"由(3.130)定义的 $\beta(\theta)$ 是否为 θ 的非降函数"这个问题. 可以证明：对一类很重要的分布，这一点成立. 特别是，其中包括了指数型分布族 $f(x,\ \theta)=C(\theta)e^{Q(\theta)T(x)}h(x)$，$Q(\theta)$ 为 θ 的严格单调函数. 这包含了例 3.14、3.15 及另一些在应用上重要的例子.

习　题

1. 把例 3.3 中的检验的否定域仔细描述出来.

2. 设 $X_1, \cdots, X_n \sim R(0,\ \theta)$，$\theta>0$. 指定 $\theta_0>0$. (a)证明 $\theta\leqslant\theta_0 \leftrightarrow \theta>\theta_0$ 的 UMP 检验存在，并算出水平 α 的 UMP 检验的功效函数. (b) 指定 $\beta\in(\alpha,\ 1)$. 证明：不论 n 多大，不存在上述检验问题的水平 α 检验，使其功效函数在对立假设上处处大于 β. (c)证明：$\theta=\theta_0 \leftrightarrow \theta\neq\theta_0$ 的 UMP 检验不存在(提示：取 $\theta_1>\theta_0$ 和 $\theta_1<\theta_0$ 分别考虑检验问题 $\theta=\theta_0 \leftrightarrow \theta=\theta_1$).

3. 设 $X_1, \cdots, X_n \sim R(\theta,\ 2\theta)$，$\theta>0$. 指定 $\theta_0>0$. 问：$\theta\leqslant\theta_0 \leftrightarrow \theta>\theta_0$ 的 UMP 检验是否存在？(提示：取 $\theta_1>\theta_0$. 考虑检验问题 $\theta=\theta_0 \leftrightarrow \theta=\theta_1$. 证明所得的水平 α 的 UMP 检验也是 $\theta\leqslant\theta_0 \leftrightarrow \theta=\theta_1$ 的 UMP 检验. 再考察此检验是否与 θ_1 有关.

4. 设 k 为已知自然数. 为检验一事件的概率 p 是否 $\leqslant p_0$，将试验独立地重复下去，直到该事件发生 k 次为止. 以 X 记到那时为止的试验次数. (a)证明 X 的分布为

$$P_p(X=x)=\binom{x-1}{k-1}p^k(1-p)^{x-k},\ x=k,\ k+1,\ \cdots.$$

(b)证明 $p\leqslant p_0\leftrightarrow p>p_0$ 的 UMP 检验存在.

5. 设 $X_1, \cdots, X_n$ 为抽自一总体的独立随机样本，总体密度为(1.8). 指定 $\lambda_0>0$. 证明 $\lambda\leqslant\lambda_0\leftrightarrow\lambda>\lambda_0$ 的 UMP 检验存在，并利用 χ^2 分布表出它.

6. (续上题)若上题中 $X_1, \cdots, X_n$ 表示受试的 n 个元件的寿命，而我们实际上只试验到有 r 个失效时即停止. 这 r 个的寿命从小到大记为 $Y_1, \cdots, Y_r$. 作出 $\theta\leqslant\theta_0\leftrightarrow\theta>\theta_0$ 的基于 $Y_1, \cdots, Y_r$ 的 UMP 检验.

7. 设 $X_1, \cdots, X_m\sim N(a, 1), Y_1, \cdots, Y_n\sim N(b, 1)$, 合样本 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ 独立, a、b 都未知. 证明: $b\leqslant a\leftrightarrow b>a$ 的 UMP 检验存在, 且有否定域 $\bar{Y}-\bar{X}>C$. (提示: 取定 $a_1<b_1$. 令 $A=(ma_1+nb_1)/(m+n)$. 考虑检验问题 $a=b=A\leftrightarrow a=a_1, b=b_1$.)

8. 设 $\varphi(x)$ 为一检验函数，而 $T=T(X)$ 为充分统计量. 证明: $E_\theta(\varphi(X)|T)$ 也是检验函数, 且其功效函数与 φ 同. 这个结果说明了什么问题?(提示: 由充分性知 $E_\theta(\varphi(X)|T)$ 与 θ 无关. 再由 φ 为检验函数证明 $E_\theta(\varphi(X)|T)$ 也是.)

9. 设 $X_1, \cdots, X_n\sim N(\theta, 1)$. 考虑检验问题 $\theta=0\leftrightarrow\theta\neq0$. 证明: 当且仅当 $d=-c(c>0)$, 以 $\{d\leqslant\bar{X}\leqslant c\}$ 为接受域的检验才是无偏的.

10. 考虑第 7 题. 试作出 $a=b\leftrightarrow a\neq b$ 的一个无偏检验.

11. 试在原假设成立的情况下，以及在对立假设的情况下，计算 Pearson χ^2 统计量(3.9)的均值.

12. 在生产同一种产品的甲、乙两厂中各抽出 n 个产品组成 n 对, 让 n 个人每人评判一对. 各人的评判是"甲好"、"乙好"、"甲乙一样"这三种之一. 试根据 n 个评判结果, 用 χ^2 检验法检验"甲、乙两厂产品质量无差别"这个假设.

13. 验证(3.64)和(3.65)式.

14. 为了检验一事件的概率 p 是否 $\leqslant1/2$($p\leqslant1/2\leftrightarrow p>1/2$). 用两种方法检验. 一种是用第 4 题的作法, 定 $\alpha=0.05$. 重复试验至该事件第 3 次不发生为止, 设到这时为止的试验次数为 12. 一种是事先预定做 12 次试验, 结果发现该事件发生 9 次. 在水平 $\alpha=0.05$ 之下, 分别用这两种情况下的试验结果去检验 $p\leqslant1/2\leftrightarrow p>1/2$(都用 UMP 检验). 二者结果不同. 本例说明了什么问题?(样本不止看它的数据, 还要看它怎么来的.)

15. 有一个二维总体 (X, Y), 其分布为:

当 $\theta=(1, 1, \cdots1)$ 时(θ 为 m 维):

$$P_\theta(X=Y=i)=P_\theta(X=Y=-i)=1/3m,\ i=1,\ \cdots,\ m,$$
$$P_\theta(X=Y=0)=1/3.$$

当 $\theta=(\theta_1,\ \cdots,\ \theta_m)\ (\theta_i\geqslant 0,\ \sum_1^m\theta_i=1)$ 时:

$$P_\theta(X=Y=i)=\theta_i/m,\ P_\theta(X=Y=-i)=0,\ i=1,\ \cdots,\ m,$$
$$P_\theta(X=Y=0)=(m-1)/m.$$

参数空间为 $\Theta=(1,\ \cdots,\ 1)\cup\{(\theta_1,\ \cdots,\ \theta_m):\theta_i\geqslant 0,\ \sum_1^m\theta_i=1\}$. 抽大小为 1 的随机样本. 作真实水平为 1/3 的似然比检验，证明其否定域为 $\{(i,\ i),\ i=1,\ \cdots,\ m\}$. 其功效函数在对立假设上之值为 $1/m$. 若 $m>3$, 则这个似然比检验还不如根本不作试验就取检验函数 $\varphi\equiv 1/3$. 试在 $m>3$ 时，找出一个真实水平为 1/3 的合理检验(此例是 E. L. Lehmann 举出的).

16. 证明: 一样本 t 检验(3.79)(单边)，确是相应假设的水平 α 检验. 对两样本 t 检验证明同一结果.

17. 在正态两样本情况下，给定 $c>0$, 找出 $\sigma_1^2=c\sigma_2^2\leftrightarrow\sigma_1^2\neq c\sigma_2^2$ 的似然比检验(提示: 通过变换把 c 化为 1).

18. 计算 § 3.7(一)中的复式抽样方案$(n_1,\ n_2,\ c_1,\ c_2,\ c)$的功效函数.

19. 设总体分布为 $R(0,\ \theta)$, 要检验假设 $\theta\leqslant 1\leftrightarrow\theta>1$. 定序贯抽样方案如下: 定自然数 m. 依次观察 $X_1,\ X_2,\ \cdots$, 若对某个 i 有 $X_i>1$ 且 $i<m$, 则停止抽样而否定 $\theta\leqslant 1$. 不然就观察到 X_m 为止，然后视 $\max(X_1,\ \cdots,\ X_m)>c$ 或否以决定是否定或接受 $\theta\leqslant 1$. 试计算此检验的功效函数及平均抽样次数(注意它们都是 θ 的函数).

20. 设总体分布为 $N(\theta,\ 1)$. 取 $\theta_0<\theta_1$. 以 φ 记 $\theta=\theta_0\leftrightarrow\theta=\theta_1$ 的序贯概率比检验. 证明: φ 的功效函数是 θ 的非降函数(提示: 任取 $\theta'<\theta''$. 考虑两个独立同分布的序列 $X_1',\ X_2'\cdots\sim N(\theta',\ 1),\ X_1'',\ X_2'',\ \cdots\sim N(\theta'',\ 1)$. 就这两个序列分别算出(3.116)式中的 S_n, 分别记为 S_n' 和 S_n''. 注意: 若记 $X_n^*=X_n'+(\theta''-\theta')$, 则由序列$\{X_n^*\}$算出的 S_n^* 与 S_n'' 同分布, 而 $S_n^*>S_n'$. 由此知: S_n^* 有更大的可能性首先从(3.117)式下面的图形的 $\log B$ 一边越出图中的带形).

第四章 区间估计

设样本 X 的分布包含未知参数 $\theta\in\Theta$，而 $g(\theta)$ 是定义在 Θ 上的一个已知函数，要利用样本 X 对 $g(\theta)$ 进行估计. 这个问题在第二章中已讨论过了. 在那里，我们是用一个由样本 X 所决定的数 $\hat{g}(x)$ 去估计 $g(\theta)$，称为 $g(\theta)$ 的点估计. 这种估计的缺点在于，单从所给出的估计值 $\hat{g}(X)$ 上，无法看出它的精度有多大. 当然，你可以给出某种指标，例如估计的均方误差之类去刻划这种精度，但这也还只是间接的. 更直接的方法，就是指出一个误差限 $d(X)$，而把估计写成 $\hat{g}(X)\pm d(X)$ 的形式. 在各种部门中我们常见到这种写法. 这事实上就是一种**区间估计**：估计 $g(\theta)$ 在区间 $[\hat{g}(X)-d(X),\ \hat{g}(X)+d(X)]$ 之内.

一般地，设有两个统计量 $A(X)$ 和 $B(X)$，满足条件 $A(X)\leqslant B(X)$，则 $[A(X),\ B(X)]$ 就可作为 $g(\theta)$ 的一个区间估计. 意思是：一旦有了样本 X，就把未知的 $g(\theta)$ 估计在区间 $[A(X),\ B(X)]$ 内.

在有些实际问题中，人们关心的只是 $g(\theta)$ 在一个方向的界限. 例如一种新材料的强度，我们关心它最低不小于多少；一个工厂的废品率，我们关心它最高不超过多少等等. 这也是一种区间估计，因为，若用 $\bar{g}(X)$ 作为 $g(\theta)$ 的上界估计，则等于说把 $g(\theta)$ 估计在区间 $(-\infty,\ \bar{g}(X)]$ 内. 同样，下界 $\underline{g}(X)$ 相当于区间估计 $[\underline{g}(X),\ \infty)$.

区间估计是一种很重要的统计推断形式. 在理论上说，它也是数理统计学中一个各家争论较多的问题. 争论的焦点在于，对区间估计问题的意义究应如何理解. 由于这种理解的不同，所引出的方法也就有差异. 就目前情况来说，占主导地位的，还是 J. Neyman 在 1934 年开始的一系列工作中所引进的置信区间理

论. 本章将对这个理论的基本概念, 以及若干重要参数的区间估计, 作一简短介绍. 本章§4.2打算介绍一个Fisher关于区间估计的观点——信任推断法. 关于用Bayes方法作区间估计的介绍则放到下一章.

§4.1 Neyman的置信区间理论

(一) 置信水平和置信系数

为书写简单计,假定被估计的$g(\theta)$就是θ本身. 一般情况没有原则区别.

设X为样本,$[\theta_1(X), \theta_2(X)]$是$\theta$的一个区间估计. 由于$\theta$未知且样本是随机的,我们不能保证在任何情况下(即对任何具体的样本值), 区间$[\theta_1(X), \theta_2(X)]$必定包含被估计的$\theta$,而只能以一定的概率保证它. 这个考虑引出下面的基本定义.

定义 4.1 如果不论参数θ在参数空间Θ中取什么值,"区间$[\theta_1(X), \theta_2(X)]$包含$\theta$"这个事件的概率, 总不小于指定的常数$1-\alpha(0\leqslant\alpha\leqslant1$, α通常很小), 即

$$P_\theta(\theta_1(X)\leqslant\theta\leqslant\theta_2(X))\geqslant1-\alpha, \text{一切 } \theta\in\Theta, \qquad (4.1)$$

则称$[\theta_1(X), \theta_2(X)]$有**置信水平**$1-\alpha$. 也常称$[\theta_1(X), \theta_2(X)]$是$\theta$的置信水平$1-\alpha$的区间估计或**置信区间**.

由此定义可知,若$1-\alpha$为置信水平,而$0\leqslant\alpha<\alpha'\leqslant1$,则$1-\alpha'$也是置信水平. 一切置信水平中的最大者称为**置信系数**. 易见, $[\theta_1(X), \theta_2(X)]$的置信系数是$\inf\{P_\theta(\theta_1(X)\leqslant\theta\leqslant\theta_2(X)):\theta\in\Theta\}$.

如果$\theta_1(X)\equiv-\infty$, 或$\theta_2(X)\equiv\infty$,则相应地得到θ的上、下界估计. 与此相应,有

定义 4.2 设$\bar\theta(X)$是θ的一个上界估计,$0\leqslant\alpha\leqslant1$. 若

$$P_\theta(\theta\leqslant\bar\theta(X))\geqslant1-\alpha, \text{一切 } \theta\in\Theta, \qquad (4.2)$$

则称$1-\alpha$为$\bar\theta(X)$的置信水平. 也常称$\bar\theta(X)$为θ的置信水平$1-\alpha$的**置信上界(或上限)**. 如前, 一切置信水平中的最大者称为

置信系数.

对下界估计情况完全类似：若

$$P_\theta(\underline{\theta}(X)\leqslant\theta)\geqslant 1-\alpha,\text{一切}\ \theta\in\Theta, \tag{4.3}$$

称 $\underline{\theta}(X)$ 为 θ 的置信水平 $1-\alpha$ 的**置信下界(或下限)**.

(4.2)式的解释是这样的：若以 $\bar{\theta}(X)$ 作为 θ 的上界，则也可能出错：即实际上有 $\theta>\bar{\theta}(X)$($\bar{\theta}(X)$非 θ 的上界)，但这种情况发生的概率不超过 α.

以上假定参数 θ 是一维的. 若 θ 是 k 维而 $k\geqslant1$，则可定义其区域估计，即一个由样本 X 决定的区域 $S(X)\subset R^k$. 意即一旦有了样本 X，就把 θ 估计在区域 $S(X)$内. $S(X)$一般有比较规则的形状，例如其各面与坐标面平行的长方体、球、椭球等等. 置信水平和置信系数的定义依旧. 例如，若 $P_\theta(\theta\in S(X))\geqslant1-\alpha$ 对一切 $\theta\in\Theta$，就称 $1-\alpha$ 是 $S(X)$的置信水平. 也说 $S(X)$是 θ 的置信水平 $1-\alpha$ 的置信区域.

置信水平和置信系数的概念是Neyman区间估计理论的基本概念. 这个概念的要点在于：被估计的参数 θ 虽然未知，但是是一个常数，没有随机性，而区间$[\theta_1(X),\ \theta_2(X)]$则是随机的. 因此，这个概念允许一种频率的解释：如果把这个区间估计反复使用许多次，则有时它包含 θ，有时不包含. 当次数充分大时，包含 θ 的频率接近于置信系数. 因此，一个置信系数为 0.95 的区间估计$[\theta_1(X),\ \theta_2(X)]$，其实际意义可理解为：当把$[\theta_1(X),\ \theta_2(X)]$使用 100 次时，平均约有 95 次，其结果是正确的，即包含了被估计的 θ.

构造置信区间的方法主要有两种. 一种是从点估计出发，一种是从假设检验出发. 在一些常见的重要问题中，这两种方法往往给出同一结果. 下面我们就来讨论这些方法.

(二) 通过点估计量构造置信区间

我们举几个例子来说明这个方法.

例 4.1 设 $X_1,\ \cdots,\ X_n$ 是抽自一总体的独立随机样本，总体

分布为指数分布(1.8). 给定 α, 找参数 λ 的置信系数 $1-\alpha$ 的置信区间和置信上、下界.

因为 $\bar{X}=\frac{1}{n}\sum_1^n X_i$ 是 $\frac{1}{\lambda}$ 的一个无偏估计(且是其 UMVUE), 我们设想 λ 的区间估计能通过 $\bar{X}$ 表出来. 因为 $\sum_1^n X_i$ 有概率密度(1.20), 由此并注意到 χ^2 分布的密度(1.28), 不难推出 $\lambda n\bar{X}\sim\chi_{2n}^2$. 于是, 若找到 a, b, $0<a<b<\infty$, 满足条件

$$P(a\leqslant\chi_{2n}^2\leqslant b)=1-\alpha, \tag{4.4}$$

则有

$$P_\lambda(a/n\bar{X}\leqslant\lambda\leqslant b/n\bar{X})=1-\alpha, \text{一切 } \lambda>0. \tag{4.5}$$

由(4.5)知, $[a/n\bar{X},\ b/n\bar{X}]$ 就是 λ 的置信系数 $1-\alpha$ 的置信区间. 它不是唯一的, 一则满足条件(4.4)的 a, b 有很多, 二则也可以设想, 不从 $\bar{X}$ 出发同样可构造 λ 的区间估计. 这样就发生一个如何在许多具同一置信系数的区间估计中选择其一的问题. 这问题到(四)中再谈, 此处可考虑选择 a, b 满足(4.4)并使 $b-a$ 最小. 这样的 a、b 也不容易决定, 实用上可行的方法是取

$$a=\chi_{2n}^2\left(1-\frac{\alpha}{2}\right),\quad b=\chi_{2n}^2\left(\frac{\alpha}{2}\right).$$

类似地, 由 $P(\chi_{2n}^2\geqslant\chi_{2n}^2(1-\alpha))=P(\chi_{2n}^2\leqslant\chi_{2n}^2(\alpha))=1-\alpha$, 分别得出 λ 的置信系数 $1-\alpha$ 的置信下、上界分别为 $\chi_{2n}^2(1-\alpha)/n\bar{X}$ 和 $\chi_{2n}^2(\alpha)/n\bar{X}$.

例 4.2 设 $X_1, \cdots, X_n\sim R(0, \theta)$, $\theta>0$. 找 θ 的置信系数 $1-\alpha$ 的置信区间和置信上、下界.

我们知道, 若记 $T=\max(X_1, \cdots, X_n)$, 则 $\frac{n+1}{n}T$ 是 θ 的 UMVUE(例 2.16). 因此, 设法去找通过 T 表出的区间估计. 因为 $X_1/\theta\sim R(0, 1)$, 由(1.25)(在其中取 $m=n$, $f(x)=1$, $F(x)=x$), 知 T/θ 有密度函数 nx^{n-1} (当 $0<x<1$, 其他处为 0). 于是, 若找 c_1, c_2, $0<c_1<c_2\leqslant1$, 满足条件

$$1-\alpha=\int_{c_1}^{c_2}nx^{n-1}dx=c_2^n-c_1^n, \tag{4.6}$$

则将有 $P_\theta(c_1 \leqslant T/\theta \leqslant c_2)=1-\alpha$, 即

$$P_\theta(T/c_2 \leqslant \theta \leqslant T/c_1)=1-\alpha,\ \theta>0. \tag{4.7}$$

于是 $[T/c_2,\ T/c_1]$ 为 θ 的一个置信系数 $1-\alpha$ 的置信区间. 要在 $0<c_1<c_2\leqslant 1$ 的范围内取 c_1、c_2, 使(4.6)成立, 并使 $\frac{1}{c_1}-\frac{1}{c_2}$ 尽可能小(以使置信区间最短). 不难证明: 这要求取 $c_1=\alpha^{1/n}$, $c_2=1$. 置信上、下界的问题留给读者.

在以上两例中, 我们依据的都是确切的分布, 因此, 作出的区间估计, 其置信系数也是确切的. 有时, 确切的分布无法找到或过于复杂不便应用. 这时 (当样本大小较大时) 可使用其极限分布, 但所得的区间估计, 其置信系数也只是近似而非确切.

例 4.3 设 $X_1, \cdots, X_n$ 是抽自具 Cauchy 分布的总体的独立随机样本, Cauchy 分布有密度函数

$$f(x,\ \theta)=\frac{1}{\pi[1+(x-\theta)^2]},\ -\infty<x<\infty,\ -\infty<\theta<\infty, \tag{4.8}$$

要作 θ 的区间估计.

以 m_n 记 $X_1, \cdots, X_n$ 的样本中位数. 因为 θ 是分布的对称中心, $m_n-\theta$ 的分布与从 $\theta=0$ 时的 Cauchy 分布中抽出的样本(大小为 n)的中位数的分布相同, 因此, 这个分布的密度函数 $f_n(x)$ 与 θ 无关. 当 n 为奇数时, $f_n(x)$ 的表达式不难利用公式(1.25)直接写出来. 当 n 为偶数时要复杂些, 但原则上并无困难. 找到 $f_n(x)$ 后, 找 $c>0$, 使

$$\int_{-c}^{c} f_n(x)dx=1-\alpha, \tag{4.9}$$

则 $P_\theta(|m_n-\theta|\leqslant c)=1-\alpha$, 对一切 θ. 由此得出 θ 的一个置信系数 $1-\alpha$ 的置信区间为 $[m_n-c,\ m_n+c]$.

但是, 因为 $f_n(x)$ 的表达式很复杂, 要由(4.9)决定 c 不容易. 但根据(1.36)式, 知当 $n\to\infty$ 时, $\sqrt{n}(m_n-\theta)$ 有极限分布 $N(0,\ \pi^2/4)$. 因此当 n 较大时, (4.9)中的 c 可近似地取为 $\pi u_{\alpha/2}/2\sqrt{n}$. 区间估计 $m_n\pm\pi u_{\alpha/2}/2\sqrt{n}$ 的置信系数近似地为 $1-\alpha$(当 $n\to\infty$

时，趋于 $1-\alpha$).

例 4.4 $X\sim B(n, p)$，找 p 的区间估计. p 是某事件的概率，而 X 是在 n 次独立重复试验中该事件发生的次数，X/n 为频率. 这个问题在历史上以"逆概率问题"的名称而知名("正问题"是指已知概率 p 去计算发生各种次数的可能性大小).

取 $T=(X-np)/\sqrt{npq}$, $q=1-p$. T 的分布仍与 p 有关，因此直接按以上几个例子处理行不通. 但当 $n\to\infty$ 时，T 的分布趋向于 $N(0, 1)$ 的分布. 因此当 n 充分大时近似地有

$$P_p(|T|\leqslant u_{\alpha/2})\approx 1-\alpha. \tag{4.10}$$

现在要把不等式 $|T|\leqslant u_{\alpha/2}$ 改写成 $c_1\leqslant p\leqslant c_2$ 的形状，c_1, c_2 与 p 无关. 若能得到这样的 c_1, c_2，则由(4.10)，有

$$P_p(c_1\leqslant p\leqslant c_2)\approx 1-\alpha. \tag{4.11}$$

于是 $[c_1, c_2]$ 可作为 p 的置信区间，置信系数近似地为 $1-\alpha$. 记 $p^*=X/n$, $q^*=1-p^*$, $u_{\alpha/2}=\lambda$. 不难解出

$$c_1,\ c_2=\frac{n}{n+\lambda^2}\left(p^*+\frac{\lambda^2}{2n}\pm\lambda\sqrt{\frac{p^*q^*}{n}+\frac{\lambda^2}{4n^2}}\right)\quad (c_1\text{ 相应负号}). \tag{4.12}$$

置信上下界也可类似求得(上界在(4.12)中取正号，下界取负号，且 λ 要改为 u_α).

(三) 通过假设检验构造置信区间

方法很简单：设要作 θ 的置信系数 $1-\alpha$ 的置信区间. 考虑检验问题

$$H: \theta=\theta_0\leftrightarrow K: \theta\neq\theta_0, \tag{4.13}$$

找出(4.13)的一个检验，它的水平为 α. 设这个检验有接受域 A_{θ_0}，即当样本 $X\in A_{\theta_0}$ 时接受 H，不然就否定 H. 则有

$$P_{\theta_0}(X\in A_{\theta_0})\geqslant 1-\alpha,\ \theta_0\in\Theta. \tag{4.14}$$

如果关系 $\{X\in A_{\theta_0}\}$ 可改写成等价的形式 $\{\theta_1(X)\leqslant\theta_0\leqslant\theta_2(X)\}$，则由(4.14)可得 $P_{\theta_0}(\theta_1(X)\leqslant\theta_0\leqslant\theta_2(X))\geqslant 1-\alpha$. 改 θ_0 为 θ，得

$$P_\theta(\theta_1(X)\leqslant\theta\leqslant\theta_2(X))\geqslant 1-\alpha,\ \theta\in\Theta. \tag{4.15}$$

(4.15)表示，$[\theta_1(X),\ \theta_2(X)]$是$\theta$的一个置信水平$1-\alpha$的区间估计. 若检验的真实水平为$\alpha$,则(4.14)可改为等号，因而(4.15)也可改为等号，故$[\theta_1(X),\ \theta_2(X)]$有置信系数$1-\alpha$.

如果要作的是θ的置信上、下限，就需要考虑单边假设$\theta\leqslant\theta_0$或者单边假设$\theta\geqslant\theta_0$的检验问题. 这些我们都将在下面通过例子说明.

例 4.5 $X_1,\ \cdots,\ X_n\sim N(a,\ \sigma^2)$, $a,\ \sigma$都未知，要求a和σ^2的置信系数$1-\alpha$的置信区间和置信上、下界.

先考虑a的问题. 在§3.6(二)中已给出了假设H: $a=a_0$的t检验，接受域为$\left\{|\sqrt{n}(\overline{X}-a_0)|/S\leqslant t_{n-1}\left(\frac{\alpha}{2}\right)\right\}$. 此处$\overline{X}$和$S^2$分别为样本均值和样本方差. 上述不等式可改写为$\overline{X}-t_{n-1}\left(\frac{\alpha}{2}\right)S\Big/\sqrt{n}\leqslant a_0\leqslant\overline{X}+t_{n-1}\left(\frac{\alpha}{2}\right)S\Big/\sqrt{n}$. 根据前面的讨论，知

$$\left[\overline{X}-t_{n-1}\left(\frac{\alpha}{2}\right)S\Big/\sqrt{n},\ \ \overline{X}+t_{n-1}\left(\frac{\alpha}{2}\right)S\Big/\sqrt{n}\right]\tag{4.16}$$

是a的一个置信系数$1-\alpha$的置信区间. 它称为**一样本t区间估计.**

若要找a的置信下界，则考虑H: $a\leqslant a_0$的检验. 在§3.6(二)中也给出了它的水平α的t检验，有接受域$\{\sqrt{n}(\overline{X}-a_0)/S\leqslant t_{n-1}(\alpha)\}$. 此不等式可改写为$\{a_0\geqslant\overline{X}-t_{n-1}(\alpha)S/\sqrt{n}\}$，于是有

$$\begin{aligned}&P_{a_0,\sigma}(a_0\geqslant\overline{X}-t_{n-1}(\alpha)S/\sqrt{n})\\=&P_{a_0,\sigma}(\sqrt{n}(\overline{X}-a_0)/S\leqslant t_{n-1}(\alpha))\\=&1-\alpha,\end{aligned}$$

此式对一切a_0成立. 改a_0为a，得出: $\overline{X}-t_{n-1}(\alpha)S/\sqrt{n}$是$a$的一个置信系数$1-\alpha$的置信下界. 同样，考虑检验问题$H$: $a\geqslant a_0$，可得a的置信系数$1-\alpha$的置信上界为$\overline{X}+t_{n-1}(\alpha)S/\sqrt{n}$.

σ^2的置信区间和置信上下界，通过考虑$\sigma^2=\sigma_0^2$、$\sigma^2\geqslant\sigma_0^2$和$\sigma^2\leqslant\sigma_0^2$的检验得到. 在§3.6(六)中，已求得这些假设的水平α的接受域分别为

$$\left\{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)\leqslant(n-1)S^2/\sigma_0^2\leqslant\chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right\},$$
$$\{(n-1)S^2/\sigma_0^2\geqslant\chi^2_{n-1}(1-\alpha)\},$$
$$\{(n-1)S^2/\sigma_0^2\leqslant\chi^2_{n-1}(\alpha)\}.$$

由此得到 σ^2 的置信系数 $1-\alpha$ 的置信区间和置信上下界分别为

$$\left[(n-1)S^2\Big/\chi^2_{n-1}\left(\frac{\alpha}{2}\right),\ (n-1)S^2\Big/\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)\right],$$
$$(n-1)S^2/\chi^2_{n-1}(1-\alpha),\ (n-1)S^2/\chi^2_{n-1}(\alpha).$$

例 4.6 $X_1, \cdots, X_m\sim N(a, \sigma^2)$, $Y_1, \cdots, Y_n\sim N(b, \sigma^2)$，且合样本 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ 相互独立. 令 $\theta=b-a$. 找 θ 的置信系数 $1-\alpha$ 的置信区间与上、下界.

在§3.6(四)中已找到了 $\theta=\theta_0$、$\theta\geqslant\theta_0$ 和 $\theta\leqslant\theta_0$ 等假设的两样本 t 检验. 记

$$S^*=\left[\frac{1}{m+n-2}\left(\sum_1^m(X_i-\overline{X})^2+\sum_1^n(Y_j-\overline{Y})^2\right)\right]^{1/2},$$

得到 θ 的置信系数 $1-\alpha$ 的置信区间和上下界分别为

$$\left[(\overline{Y}-\overline{X})-t_{m+n-2}\left(\frac{\alpha}{2}\right)\sqrt{\frac{m+n}{mn}}S^*,\right.$$
$$\left.(\overline{Y}-\overline{X})+t_{m+n-2}\left(\frac{\alpha}{2}\right)\sqrt{\frac{m+n}{mn}}S^*\right], \qquad (4.17)$$

$$(\overline{Y}-\overline{X})+t_{m+n-2}(\alpha)\sqrt{\frac{m+n}{mn}}S^*, \text{(上界)} \qquad (4.18)$$

$$(\overline{Y}-\overline{X})-t_{m+n-2}(\alpha)\sqrt{\frac{m+n}{mn}}S^*. \text{(下界)} \qquad (4.19)$$

与检验问题一样，这里我们假定了两总体有相同的方差 σ^2. 若去掉这个假设，则得到著名的 Behrens-Fisher 问题的区间估计形式(检验形式已在§3.6(四)中提到过了). 我们将在§4.2中考虑这个问题.

两正态总体方差之比的置信区间和上下界，也可以用同样的方法得到(用§3.6(六)的结果). 细节留给读者.

反过来，若我们用某种方法建立了 θ 的置信水平为 $1-\alpha$ 的区

间估计$[\theta_1(X),\ \theta_2(X)]$，则对给定的$\theta_0$，不难作出(4.13)的一个水平$\alpha$的检验．事实上，以$\{x:\ \theta_0\bar{\in}[\theta_1(x),\ \theta_2(x)]\}$为否定域的检验就是这样一个检验，证明是显然的．由此可见，区间估计和假设检验之间有很密切的关系．这种关系不止是形式上的．事实上，某种准则下的最优检验，往往导致相应准则下的最优区间估计．反之亦然．后面将更具体地谈到这一点．

与点估计和假设检验比较，区间估计这种推断形式有一个显著的特点，即它的精度(一般可用区间的长度刻划)和可靠度(用其置信系数刻划)一目了然．有如本章开始时指出的，正是因为点估计不具备这个特点，才使人们考虑区间估计，假设检验也有这个问题．考虑这样一个例子：

设从正态总体$N(a,\ \sigma^2)$中抽了一些样本去检验假设$a=0$，结果假设被接受了．如以前曾指出的，这并不意味着"证明了"$a=0$．而且，假如我们只是收到通知说$a=0$被接受了，我们甚至无法估量真正的a值与0相去能有多远．但如我们被告知：a的具一定置信系数(如0.95)的区间估计为$[-0.05,\ 0.07]$，或者是$[-15,\ 20]$，则在前一场合，a与0相距最大不过0.07，这么大小一个值在实用上可能无甚影响．这时，我们就有一定把握(0.95)说a"事实上"可认为是0，而不止是"接受$a=0$"了．若在后一场合，虽则$a=0$这个假设也被接受(因为0这个点在区间$[-15,\ 20]$内)，但因a的可能范围很大，实际上我们只能说对a"所知甚少"．

反之，若我们收到通知说"$a=0$被否定"．我们从这句话也只知道，有比较显著的证据认为$a\neq0$，但还无法知道其实际意义如何．但如我们被告知：a的区间估计为$[0.01,\ 0.03]$，或为$[-50,\ -40]$．在前一场合，虽则$a=0$被否定(因0不在区间$[0.01,\ 0.03]$内)．但a最大也不过0.03，这么小一个值可能实际上与0无异．因此，虽则在统计上否定了$a=0$，事实上可认为$a=0$．在后一场合，不仅$a=0$要否定，而且，a之值至少与0相距40，因而从实际意义上看也是显著异于0．这些分析说明：区间估计所提供的信息比假设检验更为确切些．也提醒我们：1．在实用上，对

假设检验的结果的实际含义的解释，要多小心. 2. 必要时，要参考被检验的参数的区间估计.

(四) 区间估计的优良性准则

评价一个区间估计的优劣有两个要素. 一是其**可靠度**，即区间包含未知参数的概率有多大，当然愈接近1愈好；二是其**精度**，衡量精度的一个明显指标是其长度，长度愈小愈好. 在样本大小一定的情况下，这二者是矛盾的. 若把可靠度提得极高，则区间长无限增加，而结果变得无用. 正如把一个人的年龄估在10岁到90岁之间，虽则可靠但于事无补. Neyman 的理论给定置信水平，以保证有一定的可靠度，在这个前提之下，尽量选择精度更高的区间估计.

用长度作为精度的标准不适用于置信上、下界. 就区间估计而言，以长度作标准的理论也比较复杂些. 因此又引进了另外的标准. 我们先从置信上、下界谈起.

设 $\underline{\theta}(X)$ 为 θ 的置信水平 $1-\alpha$ 的置信下界. 在满足规定的置信水平的前提下，$\underline{\theta}(X)$ 愈大愈好. 或者说，$\underline{\theta}(X)$ 愈大，它作为 θ 的下界就愈精. 正如说：甲估计 A 的年龄 $\geqslant 45$ 岁，乙估计 A 的年龄 $\geqslant 40$ 岁，则甲的估计比乙精. 指定任一个 $\theta'<\theta$. $\underline{\theta}(X)$ 愈大，θ' 落在 $[\underline{\theta}(X), \infty)$ 内的机会就愈小. 反之亦然. 因此，我们要在一定的置信水平的限制下，找这样的 $\underline{\theta}(X)$，使对任何 $\theta'<\theta$，概率 $P_\theta(\underline{\theta}(X)\leqslant\theta')$ 尽可能小. 类似地，对置信上界 $\overline{\theta}(X)$，要求当 $\theta'>\theta$ 时，$P_\theta(\overline{\theta}(X)\geqslant\theta')$ 小；对置信区间 $[\theta_1(X), \theta_2(X)]$，要求对任何 $\theta'\neq\theta$，概率 $P_\theta(\theta_1(X)\leqslant\theta'\leqslant\theta_2(X))$ 尽量小. 这些考虑引导到下面的定义：

定义 4.2 称 $\overline{\theta}^*(X)$、$\underline{\theta}^*(X)$ 和 $[\theta_1^*(X), \theta_2^*(X)]$ 分别是 θ 的置信水平 $1-\alpha$ 的**一致最精确**(UMA-Uniformly Most Accurate)**置信上、下界和置信区间**，如果它们有置信水平 $1-\alpha$，而且：

1. 对任何置信水平为 $1-\alpha$ 的置信上界 $\overline{\theta}(X)$，以及任何 $\theta<\theta'$，有 $P_\theta(\overline{\theta}^*(X)\geqslant\theta')\leqslant P_\theta(\overline{\theta}(X)\geqslant\theta')$.

2．对任何置信水平为 $1-\alpha$ 的置信下界 $\underline{\theta}(X)$，以及任何 $\theta>\theta'$，有 $P_\theta(\underline{\theta}^*(X)\leqslant\theta')\leqslant P_\theta(\underline{\theta}(X)\leqslant\theta')$.

3．对任何置信水平为 $1-\alpha$ 的置信区间 $[\theta_1(X),\ \theta_2(X)]$，以及任何 $\theta\neq\theta'$，有 $P_\theta(\theta_1^*(X)\leqslant\theta'\leqslant\theta_2^*(X))\leqslant P_\theta(\theta_1(X)\leqslant\theta'\leqslant\theta_2(X))$.

下面的定理指明由 UMP 检验构造 UMA 置信上、下界和置信区间的方法.

定理 4.1 在第(三)段中讲过的，通过由假设 $\theta=\theta_0$、$\theta\geqslant\theta_0$ 和 $\theta\leqslant\theta_0$ 去构造 θ 的置信水平 $1-\alpha$ 的置信区间和置信上、下限的方法中，若所用的检验是水平 α 的 UMP 检验，则所得的置信区间和置信上、下界是置信水平 $1-\alpha$ 的 UMA 置信区间和置信上、下界.

证 以置信下界为例，其余情况类似.

设 $\{X\in A^*(\theta_0)\}$ 为 $\theta\leqslant\theta_0\leftrightarrow\theta>\theta_0$ 的水平 α 的 UMP 检验，其产生的置信下界记为 $\underline{\theta}^*(X)$. 前已指出它有置信水平 $1-\alpha$（参看例 4.5, 4.6）. 现设 $\underline{\theta}(X)$ 为另一置信下界，有置信水平 $1-\alpha$. 引进 $\theta\leqslant\theta_0\leftrightarrow\theta>\theta_0$ 的检验，它有接受域 $\{x:\ \underline{\theta}(x)\leqslant\theta_0\}$，则其水平为 α. 事实上，若原假设成立，即 $\theta\leqslant\theta_0$，将有 P_θ(原假设被接受)$=P_\theta(\underline{\theta}(X)\leqslant\theta_0)\geqslant P_\theta(\underline{\theta}(X)\leqslant\theta)\geqslant1-\alpha$，这证明它有水平 α. 但以 $A^*(\theta_0)$ 为接受域的检验是水平 α 的 UMP 检验. 按 UMP 检验的定义，对任何 $\theta_1>\theta_0$，有

$$P_{\theta_1}(X\overline{\in}A^*(\theta_0))\geqslant P_{\theta_1}(\underline{\theta}(X)>\theta_0),$$

即 $P_{\theta_1}(X\in A^*(\theta_0))\leqslant P_{\theta_1}(\underline{\theta}(X)\leqslant\theta_0)$. 但 $\{X\in A^*(\theta_0)\}=\{\underline{\theta}^*(X)\leqslant\theta_0\}$，故 $P_{\theta_1}(\underline{\theta}^*(X)\leqslant\theta_0)\leqslant P_{\theta_1}(\underline{\theta}(X)\leqslant\theta_0)$. 此式对任何 $\theta_1>\theta_0$ 成立. 改 θ_1 为 θ，θ_0 为 θ'，得

$$P_\theta(\underline{\theta}^*(X)\leqslant\theta')\leqslant P_\theta(\underline{\theta}(X)\leqslant\theta'),\ \text{任何}\theta'<\theta.$$

这证明了 $\underline{\theta}^*(X)$ 为 UMA 置信下界. 定理证毕.

UMP 检验不常见，故 UMA 置信上、下界和置信区间也不常见. 但由定理 4.3，对指数型分布族而言，单边假设的 UMP 检验存在. 这时可以找到其 UMA 置信界. 举一个简单例子.

例 4.7 $X_1,\ \cdots,\ X_n\sim N(\theta,\ 1)$. $\theta\leqslant\theta_0\leftrightarrow\theta>\theta_0$ 的水平 α 的

UMP 检验存在，且有接受域$\{\sqrt{n}(\overline{X}-\theta_0)\leqslant u_\alpha\}$. 这相应于$\theta$的置信下界$\overline{X}-u_\alpha/\sqrt{n}$. 据定理 4.1，这是$\theta$的置信水平$1-\alpha$的 UMA 置信下界. 同理，$\overline{X}+u_\alpha/\sqrt{n}$是$\theta$的置信水平$1-\alpha$的 UMA 置信上界.

由于双边检验问题$\theta=\theta_0\leftrightarrow\theta\neq\theta_0$的 UMP 检验几乎总不存在，故 UMA 置信区间几乎总不存在，对指数型分布族参数也不例外(以此之故，定理 4.1 中关于置信区间那部分基本上是虚设的). 即使对单边检验问题，UMP 检验存在的情况也不多. 因此，对优良性准则要加以适当的放宽. 为此，引进无偏置信区间与无偏置信界的概念.

定义 4.3 $[\theta_1(X),\theta_2(X)]$、$\overline{\theta}(X)$和$\underline{\theta}(X)$分别称为$\theta$的**无偏置信区间与无偏置信上、下界**，若：

1. 对任何θ'和$\theta''\neq\theta'''$，总有

$$P_{\theta'}(\theta_1(X)\leqslant\theta'\leqslant\theta_2(X))\geqslant P_{\theta''}(\theta_1(X)\leqslant\theta'''\leqslant\theta_2(X));$$

2. 对任何θ'和$\theta''<\theta'''$，总有

$$P_{\theta'}(\overline{\theta}(X)\geqslant\theta')\geqslant P_{\theta''}(\overline{\theta}(X)\geqslant\theta''');\tag{4.20}$$

3. 对任何θ'和$\theta''>\theta'''$，总有

$$P_{\theta'}(\underline{\theta}(X)\leqslant\theta')\geqslant P_{\theta''}(\underline{\theta}(X)\leqslant\theta''').$$

总之，无偏性条件的意思是说：它包含(或$\geqslant$，或$\leqslant$)"应当包含(或应$\geqslant$，或应$\leqslant$)的值"的概率不能小于它包含(或$\geqslant$，或$\leqslant$)"不应包含(或$\geqslant$，或$\leqslant$)的值"的概率. 例如，在(4.20)的左边，θ'是参数真值，$\overline{\theta}(X)\geqslant\theta'$，$\overline{\theta}(X)$作为$\theta'$的上界估计，是应当$\geqslant\theta'$. 再看右边.$\theta''$是参数真值，$\overline{\theta}(X)$作为其上界估计，愈小愈精. 故$\overline{\theta}(X)$不应$\geqslant$比$\theta''$大的值$\theta'''$.

在这个定义的基础上，自然地就导出**一致最精确的无偏**(UMAU)**置信区间和上、下界**的概念. 它是在一切具置信水平$1-\alpha$的无偏置信区间(上、下界)中，在定义 4.2 的意义下一致最精确者.

与定理完全类似的证法不难得到下述定理：

定理 4.2 在第(三)段中讲过的，用有关检验所构造的置信

系数 $1-\alpha$ 的置信区间(上、下限)时，若所用检验为无偏的，则所得置信区间(上、下限)也无偏；若所用检验为 UMPU，则所得置信区间(上、下限)也是 UMAU.

例 4.8 设 $X_1, \cdots, X_n \sim N(a, \sigma^2)$，$a$、$\sigma$ 都未知. 由例 4.5 所构造的、a 的置信区间和置信限，都是 UMAU 的. 因为在 § 3.6 中我们曾指出(未证)，导出它们的检验——一样本 t 检验，为 UMPU 的. 同样，例 4.6 作出的两正态总体均值差 $b-a$ 的置信区间和置信限，也是 UMAU 的. 又若从 § 3.6 所给出的正态分布方差的无偏检验出发，去构造 σ 和 σ_1^2/σ_2^2 的置信区间与置信限，则都是 UMAU 的.

最后我们注意一下这个问题：前面在由 UMP 检验或 UMPU 检验去构造 UMA 或 UMAU 置信区间(或上、下界)时，我们所考虑的都是检验为非随机的情形. 若 UMP 或 UMPU 检验为随机的(如在二项分布的情形)，则问题比较复杂. 这时我们可以满足于一种近似解，也可以用一种随机化的手续表出其精确解，但这在使用上不方便.

(五) 序贯区间估计

在区间估计中使用序贯抽样的方法和动机，与假设检验的情形相同. 这里我们不打算进行广泛的讨论，只作为例子介绍一下 C. Stein 的一个重要结果.

设 $X_1, \cdots, X_n \sim N(a, \sigma^2)$，$a$、$\sigma$ 都未知，要作均值 a 的区间估计. 给定置信系数 $1-\alpha$，一个良好的区间估计是由(4.16)所给出的 t 区间估计. 这个区间估计的长为 $2St_{n-1}\left(\frac{\alpha}{2}\right)\Big/\sqrt{n}$. 如果固定 n，则因 $S=\sqrt{\sum_1^n (X_i-\bar{X})^2/(n-1)}$ 可以取任意大的值. 这区间之长不能有界. 这就是说，若事先给定一个常数 l，则不论把样本大小 n 取得多大，不能保证 t 区间的长不超过 l. 是否可能存在别的区间估计，置信系数仍为 $1-\alpha$，但具有这个性质？1940 年 Dantzig 证明了这是不可能的，他证明了：不论给定怎样小的 $1-\alpha$

>0，怎样大的 l 和 n，不存在 a 的一个基于 $X_1, \cdots, X_n$ 的区间估计，其置信系数 $\geqslant 1-\alpha$ 而区间之长总不超过 l.

如果 σ 已知，则对任给的 $1-\alpha<1$ 和 $l>0$，可找到 n 充分大，使具有上述性质的区间估计存在. 事实上，a 的区间估计 $\overline{X} \pm \sigma \times u_{\alpha/2}/\sqrt{n}$ 有置信系数 $1-\alpha$. 要使之区间估计之长 $2\sigma u_{\alpha/2}/\sqrt{n} \leqslant l$，只须取 $n \geqslant (2\sigma u_{\alpha/2}/l)^2$ 就够了. σ 愈大，所须的 n 也愈大. 由此不难理解：当 σ 未知时长度有界的置信区间（置信系数大于 0）的不存在性，就是因为 σ 可能非常大，以致任何事先指定的样本大小 n 都嫌太小.

1945 年，C. Stein 提出了一个两阶段抽样法，构造出了具有上述性质的置信区间. 他的想法在概念上很简单：在 σ 已知时，我们看到，样本大小 $n \geqslant (2\sigma u_{\alpha/2}/l)^2$ 已够了. 当 σ 未知时，我们先抽若干个样本（第一阶段抽样）以估计 σ^2，根据 σ 的估计值，看需要多大的 n 才够. 不足之数在第二阶段抽样中补齐. 现在来介绍 Stein 的方法，先证明两点预备事实.

(a) 设 m 为自然数，$\sigma^2>0$ 为常数，S^2 和 Y 都是随机变量，满足条件：1° $mS^2/\sigma^2 \sim \chi_m^2$. 2° 在给定 $S=s$ 的条件下，Y 的条件分布为 $N(0, \sigma^2/s^2)$，则 $Y \sim t_m$.

只须利用公式

$$Y\text{ 的无条件密度}=\int_0^\infty (Y\text{ 的条件密度}|S=s)(S\text{ 的密度})ds. \tag{4.21}$$

由假定 2°，被积函数第一项为 $\frac{s}{\sqrt{2\pi}\sigma}\exp\left(-\frac{s^2y^2}{2\sigma^2}\right)$. 又由 1° 及 χ_m^2 的密度形式，不难算出 S 的密度为

$$\left[2\left(\frac{m}{2}\right)^{m/2} \Big/ \left(\sigma^m \Gamma\left(\frac{m}{2}\right)\right)\right] s^{m-1} \exp\left(-\frac{ms^2}{2\sigma^2}\right)$$

（当 $s>0$，$s\leqslant 0$ 时为 0）.

以这些代入(4.21)并算出积分，不难得到：Y 的密度与自由度 m 的 t 分布密度一样.

(b) 设 $X_1, X_2, \cdots \sim N(a, \sigma^2)$，给定自然数 n_0，令

$$\overline{X}_0=\frac{1}{n_0}\sum_{i=1}^{n_0}X_i,\ S^2=\frac{1}{n_0-1}\sum_{i=1}^{n_0}(X_i-\overline{X}_0)^2. \qquad (4.22)$$

又设 $a(t)$、$b(t)$、$n(t)$都是给定于 $0<t<\infty$ 的函数, $a(t)$总不为 0, 而 $n(t)\geqslant n_0$ 取整数为值, 则有

$$Y=\frac{\sum_1^n a_i(S)(X_i-a)}{S\sqrt{\sum_1^n a_i^2(S)}}\sim t_{n_0-1}, \qquad (4.23)$$

此处 $a_i(S)=a(S)$ 当 $i\leqslant n_0$, $a_i(S)=b(S)$ 当 $i>n_0$.

证 取 $m=n_0-1$. 由定理 1.1 知 $mS^2/\sigma^2\sim\chi_m^2$, 故(a)的条件 1° 满足. 把 Y 写为

$$Y=\frac{n_0a(S)}{S\sqrt{\sum_1^n a_i^2(S)}}(\overline{X}_0-a)+\frac{b(S)}{S\sqrt{\sum_1^n a_i^2(S)}}$$

$$\times\sum_{i=n_0+1}^{n(S)}(X_i-a)=Y_1+Y_2.$$

给定 $S=s$, 先看 Y_1, 根据定理 1.1, $\overline{X}_0$ 与 S 独立. 故给定 $S=s$ 时, Y_1 的条件分布与

$$\left[n_0a(s)\Big/\left(s\sqrt{\sum_1^n a_i^2(s)}\right)\right](\overline{X}_0-a)$$

的无条件分布相同, 即为 $N(0,\ n_0\sigma^2a^2(s)\Big/\left(s^2\sum_1^n a_i^2(s)\right)$. 再看 Y_2. 因为 S 只与 $X_1,\ \cdots,\ X_{n_0}$ 有关, 故 S 与 $X_{n_0+1},\ X_{n_0+2},\ \cdots$独立, 因此, 在给定 $S=s$ 时, Y_1、Y_2 条件独立且 Y_2 的条件分布为 $N(0,(n(s)-n_0)b^2(s)\Big/\left(s^2\sum_1^n a_i^2(s)\right)$. 由此可知, 当给定 $S=s$ 时, Y 的条件分布为$N(0,\ \sigma^2/s^2)$, 故(a)的条件 2° 也成立. 利用(a)得(4.23).

现指定 $c>0$ 及自然数 n_0, 第一阶段抽样 n_0 次得 $X_1,\cdots,X_{n_0}$. $\overline{X}_0$ 和 S 的意义如(4.22). 定义

$$n(t)=\max(n_0,\ [t^2/c]+1),\ a(t)=b(t)=1/n(t). \quad (4.24)$$

若 $n(S)=n_0$, 则抽样到此为止. 若 $n(S)>n_0$, 则第二阶段再抽

$n(S)-n_0$ 次，即观察 $X_{n_0+1}, \cdots, X_{n(S)}$. 令

$$Y=\sqrt{n(S)}\,(\overline{X}-a)/S,\ \overline{X}=\frac{1}{n(S)}\sum_{i=1}^{n(S)} X_i,$$

则由以上的预备事实(b)，知 $Y\sim t_{n_0-1}$. 于是置信区间

$$\left[\overline{X}-St_{n_0-1}\left(\frac{\alpha}{2}\right)\Big/\sqrt{n(S)},\ \overline{X}+St_{n_0-1}\left(\frac{\alpha}{2}\right)\Big/\sqrt{n(S)}\right] \tag{4.25}$$

有置信系数 $1-\alpha$. 此区间之长为 $2St_{n_0-1}\left(\frac{\alpha}{2}\right)\Big/\sqrt{n(S)}$. 由(4.24)知 $n(S)\geqslant[S^2/c]+1\geqslant S^2/c$, 故区间(4.25)之长不超过 $2\sqrt{c}\times t_{n_0-1}\left(\frac{\alpha}{2}\right)$. 为使此数不超过给定的 $l>0$, 只须取 $c=l^2\Big/\left(4t^2_{n_0-1}\left(\frac{\alpha}{2}\right)\right)$ 即可.

§4.2 Fisher 的信任推断法

(一) 信任分布

差不多在 Neyman 发表其置信区间理论的同时，R. A. Fisher 提出了一种求区间估计的方法. 这个方法原则上可用于任何统计推断问题，因而它不仅是一个方法上的问题，且代表了对待统计问题的一种根本不同的观点.

Fisher 的思想可以通过一个简单例子来说明. 设有样本 $X\sim N(\theta,1)$, θ 是未知参数(样本大小为 1). 则有 $X-\theta\sim N(0,1)$, 即对任何实数 t, 有

$$P(X-\theta<t)=\Phi(t),\ \Phi(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-y^2/2}dy. \tag{4.26}$$

此式可改写为

$$P(\theta>X-t)=\Phi(t),\quad 或\quad P(\theta<X-t)=1-\Phi(t). \tag{4.27}$$

从通常的概率论观点看，由(4.26)到(4.27)，不过是写法的不同，没有什么实质性的新内容，但 Fisher 却赋予(4.27)这样一个意

义：有了样本 X 后(即 X 看成一个已知数)，把 θ 看成一个随机变量，而(4.27)给出了 θ 的分布，这分布称为 θ 的信任分布。

为什么可以把 θ 看成随机变量呢？按 Fisher 的意思，在进行抽样得到样本 X 之前，θ 是一个未知数。我们对它茫然无所知，就我们对它的了解而言，它什么值都可以取，取什么值可能性有多大，一点也说不上。但在抽样得到 X 后，虽然这时 θ 仍有可能取种种值，但对取这些值的可能性如何，通过样本 X 提供的信息，我们就有些知识了。例如，考虑到 θ 是 X 的均值，我们会觉得，θ 落在 X 附近的可能性，较远离 X 的可能性大(读者可以把这个说法与"似然性"的概念联系起来考察)。这只是一个笼统的、非定量的说法，而(4.27)则对 θ 取各种值的可能性给予定量的刻划(在得到 X 的背景下)。

因此，在这种看法之下，样本 X 的作用，或更确切地说，样本 X 所提供的信息，就表现在它破除了我们对 θ 的完全无知，而以一个概率分布的形式总结了对 θ 的新认识(有了 X 以后的认识)。

这种看法可以立即用于有关 θ 的统计推断问题。例如要作 θ 的区间估计，给定 α，我们可以找 a、b，$a<b$，使 $\widetilde{P}(a\leqslant\theta\leqslant b)=1-\alpha$($\widetilde{P}$ 表示 θ 的信任分布)。按(4.27)，这要求 a、b 满足

$$\Phi(X-a)-\Phi(X-b)=1-\alpha. \tag{4.28}$$

满足(4.28)的 a、b 很多，可以取一组 a、b，使 $b-a$ 最小。这时 $a=X-u_{\alpha/2}$，$b=X+u_{\alpha/2}$ (读者自证)。这样，我们得到 θ 的一个区间估计，其信任系数为 $1-\alpha$，此区间估计即 $[X-u_{\alpha/2}, X+u_{\alpha/2}]$。

现若从 $N(\theta, 1)$ 中抽出多于 1 个独立随机样本 $X_1, \cdots, X_n$，怎样去决定由这些样本提供的信任分布？Fisher 的作法(很自然地)是注意到 $\overline{X}$ 是 θ 的充分统计量，且 $\sqrt{n}(\overline{X}-\theta)\sim N(0, 1)$。利用这一点，像前面 $n=1$ 的情况一样，得到 θ 的信任分布是 $\theta\sim N(\overline{X}, 1/n)$。由此出发，可建立 θ 的信任系数为 $1-\alpha$ 的信任区间 $[\overline{X}-u_{\alpha/2}/\sqrt{n}, \overline{X}+u_{\alpha/2}/\sqrt{n}]$。

又如，设 $X_1, \cdots, X_n\sim N(a, \sigma^2)$，$a$ 和 σ 都是未知参数，以 $\overline{X}$ 和 S^2 分别记样本均值和样本方差，则有 $\sqrt{n}(\overline{X}-a)/S\sim t_{n-1}$。

以 $T_{n-1}(x)$ 记自由度为 $n-1$ 的 t 分布函数，有 $P(\sqrt{n}(\overline{X}-a)/S<x)=T_{n-1}(x)$. 将其改写为 $P(a>\overline{X}-Sx/\sqrt{n})=T_{n-1}(x)$, 就确定 a 的信任分布. 利用它，可得到 a 的信任系数为 $1-\alpha$ 的信任区间，即(4.16).

在以上几例中，信任分布的求得基本上是个形式上的转换，由之所决定的信任区间，与 Neyman 理论所决定的置信区间也完全一样. 因此，在 Fisher 方法发表的初期，人们认为它与 Neyman 的方法，是同一件事的两种不同的说法. 然而，往后的发展使人清楚地看到，这二者是不同的：不仅方法的解释上不同，具体结果也不同. 这一点在下面要讲的 Behrens–Fisher 问题中可以看出来.

在 Neyman 理论中，θ 是一个虽然未知，但是为非随机的常数，它谈不上有什么分布. 这个理论不需要在传统的概率论之外引进什么新概念；它在解问题(求具体参数的区间估计)时所涉及的计算和推理，都是在传统的概率论中早已确立的框框之内.

Fisher 的信任推断理论则不然，它虽则也要借用传统概率论的结果，但有两个根本问题：

1. **信任分布究竟是什么** 在基于频率解释的概率论中，分布的概念基于概率，而概率有一种可诉诸实验验证的频率解释. 在 Fisher理论中，信任分布理解为人们对 θ 取各种值的信任程度的刻划. 但这“信任程度”，一般讲可以因人而异，且无法诉诸实验验证.

这个困难还不是很实质性的. 因为，我们可以用一种类似于“公理化”的作法，对“信任分布”一词作为一个不加说明的基本概念，犹如几何公理中的一个点. 比方说，可以这样定义信任分布：它是一个由样本具体值和样本分布族所决定的一个函数，具有非降、左连续等等性质. 这当然无所不可，但这样一来就必须解决下面的问题：

2. **怎样确定信任分布** 即，要制定一些合理的规则，使用这些规则，就可以在已知样本具体值及样本分布时，唯一地定出参数的信任分布.

这个问题至今未获解决. Fisher 在充分统计量存在时，指出了如何决定信任分布，如我们前面讨论过的几个例子. 但是，充分统计量不是经常存在. 退一步说，即使存在，Fisher 也没有指出一种普遍的方法，足以唯一地决定信任分布. 如在 $X_1, \cdots, X_n \sim N(a, \sigma^2)$ 的例中，由 $\sqrt{n}(\overline{X}-a)/S \sim t_{n-1}$ 出发决定了 a 的一个信任分布. 但是，能否由别的充分统计量出发得到另外的信任分布？即使限于从 $(\overline{X}, S^2)$ 出发，能否得出另外的信任分布也不明确.

由于这些原因，Fisher 的思想在统计学界引起了相当的兴趣和争论. 大约有这样几种看法：有的学者，如 Neyman，对此是持完全否定的态度的. 有的人承认这理论不完善，但认为不妨探索一下，这些问题可否解决，或在一定范围内解决(例如 Fraser). 另一种持实用观点的人则认为，不管其基础如何，它在某些问题中提供了可用的解法，在这种情况下，不必因其基础问题未解决而拒绝不用.

(二) 用 Fisher 方法解 Behrens-Fisher 问题

前已指出：Behrens-Fisher 问题是这样一个问题：设 $X_1, \cdots, X_m \sim N(a, \sigma_1^2)$, $Y_1, \cdots, Y_n \sim N(b, \sigma_2^2)$, a、b、σ_1、σ_2 都是未知参数，且合样本 $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ 独立. 要找 $b-a$ 的区间估计(或者，要检验假设 $a=b$. 一旦解决了 $b-a$ 的区间估计问题，就可得到此假设的一个检验法，即：当区间估计不包含 0 时否定 $a=b$. 因此，下面不另讨论这个问题).

这个问题在实用上有很重要的意义. 因为在许多情况下，两总体方差相等的假定未必成立. 有不少学者在 Neyman 理论的范围内提出了一些解法，它们在这个意义下是不精确的，即其置信系数只是近似而非精确地等于指定的 $1-\alpha$. 1943 年 Scheffe 提出过一种解法，它在上述意义下是精确的(置信系数确为 $1-\alpha$). 可是这个解法有一个很大的缺点：其解依赖于样本排列的次序，而无论从那个角度看都不应有这种依赖性. 有的学者，如 Linnik，对此问题进行了很深刻的理论研究.

现在来叙述 Fisher 基于信任分布的解法．分别以 $\overline{X}$、S_1^2、$\overline{Y}$、S_2^2 记 X 样本和 Y 样本的样本均值和样本方差，以 t_1、t_2 记两个独立随机变量，其分布分别是自由度为 $m-1$ 和 $n-1$ 的 t 分布．又以 $\xi \overset{d}{=\!=} \eta$ 记随机变量 ξ 和 η 有相同分布．则有

$$\sqrt{m}(\overline{X}-a)/S_1 \overset{d}{=\!=} t_1, \quad \sqrt{n}(\overline{Y}-b)/S_2 \overset{d}{=\!=} t_2. \tag{4.29}$$

记 $Z=\overline{Y}-\overline{X}$，$\theta=b-a$．又 $S_1^*=S_1/\sqrt{m}$，$S_2^*=S_2/\sqrt{n}$．则由 (4.29) 得

$$Z-\theta=S_2^*t_2-S_1^*t_1,\ \theta=Z-(S_2^*t_2-S_1^*t_1). \tag{4.30}$$

有了样本以后，Z、S_1^*、S_2^* 可算出具体值．为醒目起见，分别把它们记为 z，s_1^* 和 s_2^*．记住它们都是已知的常数．(4.30)式给出 θ 的信任分布，即

θ 的信任分布(在已有样本，且算出 $Z=z$，$S_i^*=s_i^*$，$i=1, 2$ 时)
$=z-(s_2^*t_2-s_1^*t_1)$的通常分布． (4.31)

因为 t_1，t_2 独立且各有 t 分布，(4.31)右边不难算出．这样就确定了 θ 的信任分布，由此可以对它作出区间估计．Fisher 把它用下面的方式表出来，以便于造表：记 $r=\sqrt{s_1^{*2}+s_2^{*2}}$．并找 ψ，使 $\cos\psi=s_2^*/r$．这时 $\sin\psi=s_1^*/r$，而 $(s_2^*t_2-s_1^*t_1)=r(t_2\cos\psi-t_1\sin\psi)$．显然，$t_2\cos\psi-t_1\sin\psi$ 的(通常)分布函数只依赖于 m, n, ψ．以 $F_{m,n,\psi}$ 记这一分布函数，找 $y_{m,n,\psi,\alpha}>0$，使

$$F_{m,n,\psi}(y_{m,n,\psi,\alpha})-F_{m,n,\psi}(-y_{m,n,\psi,\alpha})=1-\alpha,$$

则(如前，$\widetilde{P}$ 记信任概率)

$$\begin{aligned}&\widetilde{P}(|z-\theta|\leqslant ry_{m,n,\psi,\alpha})\\&=P(|t_2\cos\psi-t_1\sin\psi|\leqslant y_{m,n,\psi,\alpha})=1-\alpha,\end{aligned}$$

即 $\widetilde{P}(z-ry_{m,n,\psi,\alpha}\leqslant\psi\leqslant z+ry_{m,n,\psi,\alpha})=1-\alpha$．因此，区间

$$[z-ry_{m,n,\psi,\alpha},\ z+ry_{m,n,\psi,\alpha}] \tag{4.32}$$

是 $\theta=b-a$ 的信任系数为 $1-\alpha$ 的信任区间．注意在有了样本并给定 α 后，z、r、$y_{m,n,\psi,\alpha}$ 都可以定出．Fisher 和 Yates 曾给出 $y_{m,n,\psi,\alpha}$ 的表．

Neyman 曾对这个解法提出批评．一点是有关“信任分布”本身及本例中导出信任分布的过程这方面的．由于 Neyman 和 Fisher 的立脚点不同，这一点可以姑置勿论．另一点是 Neyman 通过计算证明，若把(4.32)视为一个置信区间[1]，则其置信系数(而不是信任系数！)并非 $1-\alpha$. Neyman 指出：区间(4.32)包含被估计的 $b-a$ 的概率(不是信任概率！)，依赖于比值 $\rho=\sigma_1/\sigma_2$. 他就 $m=12$, $n=6$ 和 $\alpha=0.05$ 的情况，算出当 $\rho=0.1$, 1.0 和10时，这概率分别为 0.966, 0.960 和 0.934，而不是名义上的 0.95. Neyman 的这个批评的意义，在于明确了在本问题中，“置信系数”和“信任系数”确不是“形异实同”的东西．因而 Fisher 的方法与 Neyman 的方法不是一回事．使人感兴趣的是，Neyman 算出的值与 0.95 相去不远．因此觉得，Fisher 提供的解是可以放心使用的．

§4.3 容忍区间与容忍限

本节要讨论的问题，其提法与区间估计问题并无共同之处，但其解有形式上的相似性——都是用一个由样本确定的区间或上、下限的形式表达．因此我们把它附在这一章的末尾，作一个很简单的介绍．

（一）问题提法，容忍区间与容忍限的定义

设某工厂生产一种产品，其质量指标 X 服从正态分布 $N(a, \sigma^2)$，暂假定 a 和 σ 都已知，给定 β, $0<\beta<1$(β 通常很小)．以 $F(x; a, \sigma)$ 记 $N(a, \sigma^2)$的分布函数，可以找到许多实数 $b_1<b_2$，使

$$F(b_2; a, \sigma)-F(b_1; a, \sigma)\geqslant 1-\beta. \qquad (4.33)$$

比方说，$b_1=a-\sigma u_{\beta/2}$, $b_2=a+\sigma u_{\beta/2}$ 就是这样的一对实数．另外，也可以定出 b_3 和 b_4，分别满足以下的不等式

1) 读者谅必注意到，置信区间和信任区间在外形上并无区别，它们都是一个依赖于(且只依赖于)样本的区间．不同之处在于得出的过程及意义的解释．

$$F(b_3; a, \sigma) \leqslant \beta, \ F(b_4; a, \sigma) \geqslant 1-\beta. \qquad (4.34)$$

比方说，可取

$$b_3 = a - \sigma u_\beta, \ b_4 = a + \sigma u_\beta.$$

如果一对实数b_1, b_2满足(4.33)，则在该工厂生产的产品中，至少有$100(1-\beta)\%$，其质量指标落在b_1, b_2之间．同样，若b_3, b_4满足(4.34)，则产品中至少有$100(1-\beta)\%$，其指标不小于b_3，也有$100(1-\beta)\%$，其指标不大于b_4．这类知识可能有重大的实际意义．比方说，若指标愈大产品质量愈好，且甲等品要求指标值不小于指定的数B．给定$\beta=0.01$，求出b_3．若$b_3 \geqslant B$，则表示在产品中至少有99%的甲级品．又如，指定了两个数$B_1 < B_2$．只有当指标值落在B_1、B_2之间时，产品才合格．指定$\beta=0.01$，要求产品的合格率至少为99%．这时问题归结为：能否找到b_1、b_2，使(4.33)成立，且$B_1 \leqslant b_1 < b_2 \leqslant B_2$．

如果a和σ确实已知因而X的分布已知，这里就不存在什么统计问题，只须查正态分布表算一下就行．如果a、σ未知，则满足(4.33)和(4.34)的b_1, b_2, b_3, b_4需要由X的观察值去估计（比方说，b_1、b_2可以考虑用$\overline{X} \pm S u_{\alpha/2}$去估计，$S^2$为样本方差）．设想用某个估计量$\hat{b}_i(X_1, \cdots, X_n) = \hat{b}_i$去估计$b_i$．这时产生一个问题：$F(\hat{b}_2; a, \sigma) - F(\hat{b}_1; a, \sigma)$是否能$\geqslant 1-\beta$？由于$\hat{b}_1, \hat{b}_2$有随机性，我们显然不能绝对保证这一点．于是就只好降低要求：给定γ，$0 < \gamma < 1$（γ通常很小），要求“$F(\hat{b}_2; a, \sigma) - F(\hat{b}_1; a, \sigma) \geqslant 1-\beta$”这个关系至少以概率$1-\gamma$成立．对$\hat{b}_3, \hat{b}_4$也可提出类似的要求．这就引导到容忍区间和容忍限的概念．以上我们假定变量X的分布是正态，这一点当然不是本质的，对别的分布也可提出同样的问题．

因此，现设变量X有分布$F(x)$．分布F完全未知，或者其形状已知但有些参数未知．$X_1, \cdots, X_n$是X的独立随机样本．设给定了β, γ，$0 < \beta < 1$，$0 < \gamma < 1$．又设$T_i = T_i(X_1, \cdots, X_n)$，$i = 1、2、3、4$，都是统计量，满足条件$T_1 \leqslant T_2$．

定义 4.4 称$[T_1, T_2]$为F的一个(β, γ)容忍区间，若

$$P_F(F(T_2)-F(T_1)\geqslant 1-\beta)\geqslant 1-\gamma \tag{4.35}$$

称 T_3、T_4 分别是 F 的 $(\beta,\ \gamma)$ **容忍下、上限**，若

$$P_F(1-F(T_3)\geqslant 1-\beta)=P_F(F(T_3)\leqslant\beta)\geqslant 1-\gamma; \tag{4.36}$$

$$P_F(F(T_4)\geqslant 1-\beta)\geqslant 1-\gamma. \tag{4.37}$$

此处 P_F 表示：在计算事件$\{F(T_2(X_1,\ \cdots,\ X_n))-F(T_1(X_1,\cdots,X_n))\geqslant 1-\beta\}$等的概率时，$X_1,\ \cdots,\ X_n$ 各有分布 F.

我们注意到，从形式上看，容忍区间和容忍限与置信区间和置信限有相似之处，但实质上是不同的东西．主要在于：后者的目的是估计分布中的未知参数，而前者则并非如此．拿容忍区间来说，因为满足条件 $F(b_2)-F(b_1)\geqslant 1-\beta$ 的实数对有无穷多，因此不能说 T_1、T_2 是为估计 b_1、b_2．即使退一步说，我们用某种方法选择一组确定的 b_1、b_2（如正态分布中的 $b_1,\ b_2=a\pm\sigma u_{\beta/2}$），我们仍不能说 T_1、T_2 是为估计 $b_1,\ b_2$．因为在容忍区间的定义中，我们并不关心 T_1、T_2 是否与 b_1、b_2 接近，而是关心 $F(T_2)-F(T_1)$是否 $\geqslant 1-\beta$．这当然不是一回事．

话虽如此说，这二者之间还是存在一定的联系．例如，仿照定义 4.2，可定义$(\beta,\ \gamma)$**UMA 容忍限**，且可以证明：UMA 置信界和 UMA 容忍限之间有一定的联系．有关细节这里都不谈了，我们只就几种情况来讨论一下求容忍区间和容忍限的方法．

（二）X 的分布函数 $F(x)$完全未知，但假定它在$(-\infty,\ \infty)$处处连续

这是一个典型的非参数统计问题．其解法是次序统计量的一项应用．先证明一个预备事实．

引理 4.1 设一维变量 $X\sim F(x)$且 $F(x)$处处连续，则 $Y=F(X)\sim R(0,\ 1)$.

证 因为 $0<Y<1$，只须对 $0<y<1$ 证明 $P(Y<y)=y$，记 $t=\inf\{x:\ F(x)\geqslant y\}$，则由 $F(x)$处处连续且非降，易见 $F(t)=y$，以及 $F(x)<y\Leftrightarrow x<t$．故

$$P(Y<y)=P(F(X)<y)=P(X<t)=F(t)=y.$$

引理证毕.

现设 $X_1, \cdots, X_n$ 为 X 的独立观察值，$X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 为其次序统计量. 根据引理 4.1 知,若记 $U_i = F(X_i)$, $i=1, \cdots, n$,则 $U_1, \cdots, U_n \sim R(0, 1)$. 因此，若以 $U_{(1)} \leqslant \cdots \leqslant U_{(n)}$ 记 $U_1, \cdots, U_n$ 的次序统计量,则 $U_{(i)} = F(X_{(i)})$, $i=1, \cdots, n$. 记 $V_{ij} = U_{(j)} - U_{(i)}$, $1 \leqslant i < j \leqslant n$, V_{ij} 的密度已在(1.27)式给出,有

$$P(F(X_{(j)}) - F(X_{(i)}) \geqslant 1-\beta) = P(V_{ij} \geqslant 1-\beta)$$
$$= \int_{1-\beta}^{1} g_{nij}(v) dv. \qquad (4.38)$$

其中 g_{nij}由(1.27)式给出。如果选择 i, j, 使(4.38)式中的积分不小于给定的 $1-\gamma$,则根据定义,$[X_{(i)}, X_{(j)}]$就是 F 的一个(β, γ)容忍区间. 一般,尽可能选择 i、j, 使 $i+j=n+1$,或 n, $n+2$ 也可以. 这样得出的区间一般较短些. 若对某个 n, 即使取 $i=1$ 和 $j=n$, (4.38)的积分也比 $1-\gamma$ 小,则这个方法行不通. 这时,或者降低 $1-\beta$ 或 $1-\gamma$ 之值,或增大 n。

一个分布若有形如(1.27)的密度,就称为 **Beta 分布**,参数为 $j-i$ 和 $n-j+i+1$,记为 $\mathrm{Be}(j-i, n-j+i+1)$. Beta 分布的参数不必为整数,只要大于 0 就行: 当 $p>0$, $q>0$ 时,$\mathrm{Be}(p, q)$表示一分布,分布函数为

$$I_{p,q}(x) = \frac{1}{B(p, q)} \int_0^x t^{p-1}(1-t)^{q-1} dt. \qquad (4.39)$$

$B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt$, 称为 **Beta 积分**. 熟知它与 Gamma 函数 $\Gamma(x)$有关系: $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$. (4.39)当 $0<x<1$ 时称为**不完全 Beta 积分**, Pearson 曾给它造了表. 这表可用于选择 i, j 的问题.

F 的容忍上下限的问题也同样处理. 有

$$P(F(X_{(i)}) \geqslant 1-\beta) = P(U_{(i)} \geqslant 1-\beta).$$

利用(1.25)式,在其中置 $F(x)=x$, $f(x)=1$, 得

$$P(F(X_{(i)}) \geqslant 1-\beta) = \int_{1-\beta}^{1} g_{n,i,n-i+1}(v) dv, \qquad (4.40)$$

$$P(F(X_{(j)})\leqslant\beta)=\int_0^\beta g_{n,j,n-j+1}(v)dv. \qquad (4.41)$$

选择 i, j, 使(4.40)和(4.41)的右边的积分 $\geqslant 1-\gamma$, 则 $X_{(i)}$ 是 F 的(β, γ)容忍上限, 而 $X_{(j)}$ 是 F 的(β, γ)容忍下限, 它们可借助于不完全 Beta 函数表求得.

(三) 正态分布的容忍区间与容忍限

实用上有重要意义的是找正态分布的容忍区间与容忍限. 这个问题在本节一开始就提出来了.

因为正态分布函数处处连续, 所以上一段的解法适用于它. 但是这一解法没有利用正态分布的特点, 因此结果比较粗糙. 这里我们用基于充分统计量($\overline{X}$, S^2)的解, 它在直观上更容易理解.

设有样本 $X_1, \cdots, X_n \sim N(a, \sigma^2)$, a 和 σ 都未知. 在本节开始处已指出, 若 a、σ 已知, 则(β, γ)容忍上下限和容忍区间分别为 $a+\sigma u_\beta$、$a-\sigma u_\beta$ 和 $[a-\sigma u_{\beta/2}, a+\sigma u_{\beta/2}]$. 现 a、σ 未知, 可分别用 $\overline{X}$ 和 S 去估计它. 由于估计而带来的随机性, (β, γ)容忍上限不见得正是 $\overline{X}+Su_{\beta/2}$, 而可能要将系数 $u_{\beta/2}$ 修改为某个 λ, λ 既与 β 有关, 也与 γ 有关(注意 $u_{\beta/2}$ 与 γ 无关. 仅由这一点就可知道 $\overline{X}+Su_{\beta/2}$ 非(β, γ)容忍上限). 容忍下限和容忍区间也如此.

因此, 我们来找 λ, 使 $\overline{X}+\lambda S$ 为(β, γ)容忍上限, 以 $F_{a,\sigma}(x)$ 记 $N(a, \sigma^2)$ 的分布函数. 为要 $F_{a,\sigma}(\overline{X}+\lambda S)\geqslant 1-\beta$, 充要条件是 $\overline{X}+\lambda S\geqslant a+\sigma u_\beta$. 于是问题归结为求 λ, 使

$$P(\overline{X}+\lambda S\geqslant a+\sigma u_\beta)\geqslant 1-\gamma.$$

记 $S'=S/\sigma$, $\sqrt{n}(\overline{X}-a)/\sigma=Y$, 由上式得

$$P\left(\frac{Y-\sqrt{n}\,u_\beta}{S'}\geqslant -\sqrt{n}\,\lambda\right)\geqslant 1-\gamma. \qquad (4.42)$$

因为 Y、S' 独立, $S'\sim\sqrt{\frac{1}{n-1}\chi^2_{n-1}}$, $Y\sim N(0, 1)$, 故由非中心 t 分布的定义, 知 $(Y-\sqrt{n}u_\beta)/S'\sim t_{n-1,\delta}$, $\delta=-\sqrt{n}\,u_\beta$. 如果能

够确定 λ'，使 $P(t_{n-1,\delta}\geqslant\lambda')=1-\gamma$，则 $\lambda=-\lambda'/\sqrt{n}$ 满足(4.42)，而 (β,γ) 容忍上限为 $\overline{X}+\lambda S$，由正态分布的对称性，易知 (β,γ) 容忍下限为 $\overline{X}-\lambda S$。当然，现有的非中心 t 分布表规模还不够大，不一定能直接从表上查出 λ'。

(β,γ) 容忍区间可以利用上面求得的容忍上、下限，再利用下述的一般规则定出：

引理 4.2 若 T_1，T_2 分别是 F 的 $(\beta/2,\gamma/2)$ 的容忍上、下限，且总有 $T_1\geqslant T_2$，则 $[T_1, T_2]$ 为 F 的 (β,γ) 容忍区间。

证明很容易，留给读者作为练习。但用这种方法作出的 (β,γ) 容忍区间一般失之过粗。Wald 和 Wolfowitz 提出了一种近似解法如下：以 $\Phi(x)$ 记 $N(0,1)$ 的分布函数，找 b，使 $\Phi\left(\frac{1}{\sqrt{n}}+b\right)-\Phi\left(\frac{1}{\sqrt{n}}-b\right)=1-\beta$。再算出 $\lambda=\sqrt{n-1}\,b/\sqrt{\chi^2_{n-1}(\gamma)}$。则 $[\overline{X}-\lambda S, \overline{X}+\lambda S]$ 近似地为 F 的 (β,γ) 容忍区间。这解法的一个优点是计算容易些，且 λ 有表可查。

习　题

1. 设 $X_1,\cdots,X_m\sim N(a,\sigma_1^2)$，$Y_1,\cdots,Y_n\sim N(ca,\sigma_2^2)$，$c$、$\sigma_1$、$\sigma_2$ 已知，a 未知，$c\neq0$。又合样本 $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 独立。(a)找 a 的 UMVUE。(b)基于此 UMVUE，构造 a 的一个置信系数为 $1-\alpha$ 的置信区间(提示：决定常数 d，使 $d\overline{X}+(1-d)\overline{Y}/c$ 的方差最小。证明这是 a 的无偏估计且达到 c－R 不等式的下界)。

2. 设 $X_1,\cdots,X_n\sim R(\theta_1,\theta_2)$，$-\infty<\theta_1<\theta_2<\infty$，$\theta_1$，$\theta_2$ 都未知。(a)找出 $\theta_2-\theta_1$ 的一个置信系数为 $1-\alpha$ 的置信区间(提示：证明 $[(\theta_2-\theta_1)-(\max\limits_{1\leqslant i\leqslant n}X_i-\min\limits_{1\leqslant i\leqslant n}X_i)]/(\theta_2-\theta_1)$ 的分布与 θ_1，θ_2 无关)。(b)找出 $(\theta_1+\theta_2)/2$ 的置信系数为 $1-\alpha$ 的置信区间(提示：证明

$$(\max_{1\leqslant i\leqslant n}X_i+\min_{1\leqslant i\leqslant n}X_i-(\theta_1+\theta_2))/(\max_{1\leqslant i\leqslant n}X_i-\min_{1\leqslant i\leqslant n}X_i)$$

的分布与 θ_1，θ_2 无关。为此不失普遍性可设 $\theta_1=0$)。

3. 设 $X_1,\cdots,X_m\sim R(0,\theta_1)$，$Y_1,\cdots,Y_n\sim R(0,\theta_2)$，$\theta_1>0$，$\theta_2>0$ 都未知，且合样本独立。找出 θ_1/θ_2 的一个置信系数为 $1-\alpha$ 的置信区间(提示：

令 $T_1=\max\limits_{1\le i\le m} X_i$, $T_2=\max\limits_{1\le i\le n} Y_i$. 证明 $\frac{T_2}{T_1}\frac{\theta_1}{\theta_2}$ 的分布与 θ_1、θ_2 都无关).

4. 具有密度

$$f(x,\ \theta_1,\ \theta_2)=\begin{cases}\frac{\theta_2}{\theta_1}\left(\frac{x}{\theta_1}\right)^{\theta_2-1}e^{-(x/\theta_1)^{\theta_2}}, & x>0;\\ 0, & x\leqslant 0\end{cases}$$

的分布称为 **Weibull** 分布, $\theta_1>0$, $\theta_2>0$ 都是未知参数. 设 $X_1, \cdots, X_n$ 为抽自具此分布的总体的独立随机样本. (a)写出为估计 θ_1, θ_2 的似然方程. (b)令 $Y_i=(X_i/\theta_1)^{\theta_2}$, $i=1, \cdots, n$, 证明 $Y_1, \cdots, Y_n$ 独立同分布, 且 Y_1 的分布与 θ_1, θ_2 无关. (c)将似然方程转化到 $Y_1, \cdots, Y_n$, 以证明若 $\hat{\theta}_2$ 为似然方程对 θ_2 之解, 则 $\hat{\theta}_2/\theta_2$ 的分布与 θ_1、θ_2 都无关. (d) 指出利用(c)中之结果去构造 θ_2 的置信区间的方法(此方法有表可查).

5. 设 $X_1, \cdots, X_n$ 为从具密度

$$f(x,\ \theta)=\begin{cases}e^{-(x-\theta)}, & x>\theta;\\ 0, & x\leqslant\theta\end{cases}$$

的总体中抽出的独立随机样本, $-\infty<\theta<\infty$, θ未知. 证明: $\min\limits_{1\le i\le n} X_i-\theta$ 的分布与 θ 无关. 求出此分布, 从而确定 θ 的置信系数为 $1-\alpha$ 的置信区间(提示: 记 $X'_i=X_i-\theta$, $i=1, \cdots, n$. 证明 $X'_1, \cdots, X'_n$ 的分布与 θ 无关, 又注意 $\min\limits_{1\le i\le n} X_i-\theta=\min\limits_{1\le i\le n} X'_i$).

6. 设 $X\sim B(m,\ p_1)$, $Y\sim B(n,\ p_2)$. m、n 很大, p_1、p_2 是未知参数. 作出基于 X、Y 的、p_2-p_1 的置信区间, 渐近置信系数为 $1-\alpha$.

7. 设 $X_1, \cdots, X_m$ 和 $Y_1, \cdots, Y_n$ 分别为自具参数 θ_1 和 θ_2 的 Cauchy 分布总体中抽出的独立随机样本(参数为 θ 的 Cauchy 分布有密度 $\{\pi[1+(x-\theta)^2]\}^{-1}$), 合样本独立且 m、n 很大. 作 $\theta_2-\theta_1$ 的置信区间, 渐近置信系数为 $1-\alpha$(提示: 利用样本中位数及其极限定理).

8. 设 $X_1, \cdots, X_m$ 和 $Y_1, \cdots, Y_n$ 分别是从具参数 λ_1 和 λ_2 的指数分布中抽出的独立随机样本(参数为 λ 的指数分布有密度 $\lambda e^{-\lambda x}$ 当 $x>0$ ($x<0$ 时为0)), 合样本独立. 试确定 $\lambda_2-\lambda_1$ 的一个信任系数为 $1-\alpha$ 的信任区间(不必算到底, 指出方法就可以. 提示: 利用 $\lambda_1 m\overline{X}\sim\chi^2_{2m}$, $\lambda_2 n\overline{Y}\sim\chi^2_{2n}$).

9. 设 $X_1, \cdots, X_m$ 和 $Y_1, \cdots, Y_n$ 分别是从两个总体中抽出的独立随机样本, 且合样本独立. 仿上题的方法求: (a)第一、二总体的分布分别为 $N(a,\ \sigma_1^2)$ 和 $N(b,\ \sigma_2^2)$, a、b、σ_1^2、σ_2^2 都未知, 求 $\sigma_2^2-\sigma_1^2$ 的信任系数为 $1-\alpha$ 的信任区间. (b) 第一、二总体的分布分别是 $R(0,\ \theta_1)$ 和 $R(0,\ \theta_2)$, $\theta_1>0$, $\theta_2>0$ 未知, 求 $\theta_2-\theta_1$ 的信任系数为 $1-\alpha$ 的信任区间.

10. 设 $X_1, \cdots, X_n$ 为自指数分布(1.8)中抽出的独立随机样本, 定出

λ 的信任分布，由之定出其信任系数为 $1-\alpha$ 的最短信任区间(提示：用充分统计量 $\overline{X}$).

11. 设样本与总体分布同上题. 试定出形如 $C/\overline{X}$ 的 (β, γ) 容忍上下限，并利用引理 4.2, 定出 (β, γ) 容忍区间.

12. 设 $X_1, \cdots, X_n \sim R(0, \theta)$, $\theta>0$ 未知. 试利用统计量 $\max\limits_{1\leqslant i\leqslant n} X_i$, 定出 (β, γ) 置信上、下限.

第五章 Bayes 统计与统计判决理论

学过初步概率论的人都知道 Bayes 公式. 此公式包含在英国学者 T.Bayes 于 1763 年(在他去世后两年)发表的一篇文章中. 从形式上看,它不过是条件概率定义的一个简单推论,但却包含了归纳推理的一种思想,这一点在 Bayes 文章题目中已点明了. 后世的学者把它发展为一种关于统计推断的系统的理论和方法,通称 Bayes 统计. 信奉 Bayes 统计,乃至鼓吹 Bayes 观点是关于统计推断的唯一正确观点的那些统计学者,组成数理统计学中的 Bayes 学派.

在本世纪二、三十年代,就有一些学者,例如 Jeffreys, Keynes,鼓吹 Bayes 学派的观点. 但 Bayes 学派取得较大的影响,还要算到战后特别是六十年代以来的事情. 时至今日,每个学习数理统计学的人,都应当对这个学派的观点和方法有所了解. 对主要兴趣在于应用者也不能例外.

Bayes 学派的观点在统计学界引起了热烈的争论. 在介绍 Bayes 统计的基本内容时,不能也不应回避这些问题. 有一点需要说明的是:虽然本书在讨论这种问题时,难免要渗杂作者自己的一些看法,但我们将力求做到从不偏不倚的立场去介绍争论中的种种观点,希望使读者正确地理解问题实质之所在,以形成自己的看法.

统计判决理论是 A. Wald 在 1950 年的著作«Statistical Decision Function»中提出的一种统计理论. 这个理论对战后时期数理统计学的发展起了较大的影响,可以认为,现在它已成了数理统计基础结构中的一个组成部分,在应用上也有其重要性. 它与 Bayes 统计并无逻辑上的包含或承前启后性质的关系. 二者的联系,一方面在于 Bayes 判决是统计判决中的一个重要部分,

一方面在于 Bayes 方法在整个统计判决理论中是一个重要工具.因此我们把这两个主题放在这同一章中来讨论.

在本章中,我们不可能对这两个主题作详细的阐述.我们的目的,除了介绍一些有用的方法以外,尤其重要的是,希望读者能对所论问题有一个虽然只是初步、但还是清楚而比较系统的了解.没有这一点了解,就谈不上对现代数理统计学有一个较全面的认识.

§5.1 Bayes统计推断

(一) 从 Bayes 公式谈起

前面提到,Bayes 统计的基本观点是由 Bayes 公式引伸而来.我们知道 Bayes 公式,也了解什么是统计推断,因此我们有条件考察一下,这 Bayes 公式是怎样与统计推断挂上钩的.

现设有金银铜三种合子,其中金合 5 个,银合 4 个,铜合三个.每个合子里放了红黄蓝白四种球,个数为:金合:红 70,黄 20,蓝 8,白 2;银合:红 10,黄 75,蓝 3,白 12;铜合:红 5,黄 12,蓝 80,白 3.在这 12 个合子中随机抽取一个(每个被抽到的概率为 $\frac{1}{12}$),再从这合子里随机抽取一个球(每个被抽到的概率为 $\frac{1}{100}$),发现结果是红的.问:"此球是从一金合中抽出"这事件的概率是多少?

若把这概率记为 $P(金|红)$,则 Bayes 公式给出

$$P(金|红)=\frac{5}{12}\cdot\frac{70}{100}\Big/\left(\frac{5}{12}\cdot\frac{70}{100}+\frac{4}{12}\cdot\frac{10}{100}+\frac{3}{12}\cdot\frac{5}{100}\right)=70/81.$$

各种可能情况计算结果如下:

$$\begin{array}{lll}P(金|红)=70/81, & P(银|红)=8/81, & P(铜|红)=3/81,\\ P(金|黄)=25/109, & P(银|黄)=75/109, & P(铜|黄)=9/109,\\ P(金|蓝)=10/73, & P(银|蓝)=3/73, & P(铜|蓝)=60/73,\\ P(金|白)=10/55, & P(银|白)=36/55, & P(铜|白)=9/55.\end{array}\qquad(5.1)$$

从概率计算上看这是一道最普通不过的题，可是我们可以改用一种有点统计气味的提法："要由抽出的球的颜色，去推断它所来自的盒子的材料". (5.1)这张表，从一个意义上说，给出了推断结果，因为它给出了各种情况可能性的大小，它表示我们知识所达到的限度. 如果一定要给出一个判然的回答，可以取可能性最大者，观表5.1知应为:

$$红\to金;\ 黄\to银;\ 蓝\to铜;\ 白\to银. \tag{5.2}$$

(5.2)是一个完整的统计推断程序，它告诉我们，在各种结果之下问题该如何回答.

我们知道：统计推断问题是有了样本 X，X 的分布族 $\{F_\theta(x),\ \theta\in\Theta\}$，$\Theta$ 为参数空间，要由 X 去推断 θ. 上面这个例子完全可纳入这个模式. 为此引进 X, θ:

$$\theta(金)=1,\ \theta(银)=2,\ \theta(铜)=3;\ 参数空间\Theta=\{1,\ 2,\ 3\}.$$

$$X(红)=1,\ X(黄)=2,\ X(蓝)=3,\ X(白)=4.$$

$$样本空间: \mathscr{X}=\{1,\ 2,\ 3,\ 4\}.$$

而样本分布族为

$F_1(1)=70/100, F_1(2)=20/100, F_1(3)=8/100,\ F_1(4)=2/100;$

$F_2(1)=10/100, F_2(2)=75/100, F_2(3)=3/100, F_2(4)=12/100;$

$F_3(1)=5/100, F_3(2)=12/100, F_3(3)=80/100,\ F_3(4)=3/100.$

统计推断问题为由样本 X(球的颜色)推断 θ(盒子的材料). 在这个提法之下，可以把(5.2)看成一个估计量 $\hat{\theta}$: $\hat{\theta}(1)=1$, $\hat{\theta}(2)=\hat{\theta}(4)=2$, $\hat{\theta}(3)=3$.

但是，我们应注意到，这个例子的情况与前几章所论统计推断问题相比，有一个不同的地方. *即在本例中我们有(5.1)式，此式是从一些假定出发经严密推导而得，并不依赖任何特殊的统计推断方法*. 这(5.1)式比一个具体的推断方法——例如(5.2)——更为基本和重要. 因为它告诉我们，在种种具体的样本之下我们对参数 θ 能了解到何种程度，它不取决于所用统计推断方法. 相反，由它可以产生种种推断方法，例如(5.2).

为什么在前几章中讨论过的统计推断问题中，没有出现象(5.1)这种性质的结果呢？这是因为，在此例中被推断的 θ 本身就是一随机变量(抽出那种材料的合子是随机的)，且我们知道了 θ 的分布，即

$$P(\theta=1)=5/12,\ P(\theta=2)=4/12,\ P(\theta=3)=3/12. \tag{5.3}$$

这一点正是本例异于以前所处理过的种种统计推断问题的决定性因素，也是 Bayes 统计区别于前几章那种统计的特征所在。下面将细谈这一点。

（二）先验分布与后验分布

设样本 X 有分布族$\{F_\theta(x),\ \theta\in\Theta\}$，要由 X 推断 θ。这里，模型，即 X 的分布族，提供了一种知识。这个知识是对所研究的事物的知识，不是关于 θ 的知识，但对推断 θ 有用。另一种知识由样本 X 提供，它包含了有关 θ 的信息。在前几章，为推断 θ，就依靠这两种知识。Bayes 统计与此不同之处，在于它还需要另外一个前提：要预先给出 θ 取各种可能值的概率。形式上说，它要求把 θ 看成一个随机变量，并给定 θ 的概率分布 $H(\theta)$。由于这个分布 $H(\theta)$是在抽样(观察 X)以前就给出来了，故把它称为 θ 的**先验分布**。所谓"先验"，只不过是指在抽样之先，并无其他含义。

定义5.1 参数空间 Θ 上的任一概率分布，叫 θ 的先验分布。

在有些情况下，θ 的先验分布存在是一个合理的假定。（一）中那个例子就是如此，且先验分布即为(5.3)。另一个有实用意义的例子如下：某厂每天在当天成品中抽样估计其废品率 p。从当天看，p 是一单纯的未知数。但从较长一个时期看，每天都有一个 p 值，其值因随机因素的作用逐日有些波动。本日的 p 值可合理地视为随机变量 p 的一个可能值。如果我们有相当长一个时期的逐日废品率记录，则可以相当精确地定出 p 的先验分布。

在另一些情况下，把参数 θ 作为一个随机变量这种看法是勉强的。例如，要估计的 θ 是一个铁矿的矿石含铁百分率。这时，要

把 θ 看成随机变量，就要设想这铁矿是无穷多"类似"铁矿的一个样品，这是不自然的. 在此，干净利落的看法就是把 θ 作为一个孤立的未知常数. 在另一些情况下，虽则把 θ 作为一个随机变量有一定的理由，但人们对它的先验知识没有确切到能以一个分布表出的程度. 例如，我们可能知道，某工厂的废品率 θ 经常在 0.01 附近波动，偶尔也取很接近 0 或 1 之值，但我们的识知还不够确切到可以写出 θ 的先验分布. 按照 Bayes 学派的主张，即使在这种情况下，也必须提出一个分布作为 θ 的先验分布. 这可以是通过大致的估计，或掺杂着主观设想的成份以至全然是主观的，或是数学上的一种方便形式(如所谓共轭先验分布，见本节(五))，什么都可以.

既已有了 θ 的分布(先验分布)，以及给定 θ 的条件下 X 的条件分布(样本分布)，就可以定下随机变量(θ, X)的联合概率分布. 在这个联合分布之下 X 的分布，称为 X 的边缘分布. 不要把后者与 X 的样本分布混为一谈：样本分布与 θ 有关. 边缘分布是样本分布对 θ 的"平均"(按先验分布的概率去平均)，它是与 θ 无关的. 因此，Bayes 统计推断问题可以提成：有一个随机变量 (θ, X)，其联合分布完全已知，但 θ 不能观察而只能观测 X. 要由 X 去推断 θ.

现在来引进在 Bayes 统计中极重要的**后验分布**的概念.

定义 5.2 在得到样本 $X=x$ 后，θ 的后验分布，就是在给定 $X=x$ 的条件下，θ 的条件分布.

后验分布既与 x 有关，也与先验分布 H 及 X 的样本分布族有关，它是按通常概率论中计算条件分布的公式去确定的. 例如，设 X 的分布有概率密度 $f(x, \theta)$，而 θ 的先验分布 H 有密度函数 $h(\theta)$，则由概率论中计算条件密度的公式，知 θ 的条件密度，即后验密度，为

$$h(\theta|x)=f(x, \theta)h(\theta)\Big/\int_{\Theta}f(x, \varphi)h(\varphi)d\varphi. \qquad (5.4)$$

(5.4)右边分母只与 x 有关而与 θ 无关. 由于这个原因，有时分母

之值没有必要计算出来，观以后的例子可知. 现在举几个简单例子.

例 5.1 设 $X\sim$ 二项分布 $B(n, p)$. 给定 p 的先验分布为 $(0, 1)$ 均匀分布 $R(0, 1)$. 由(5.4)，知 θ 的后验密度为(当样本 $X=x$ 时)

$$h(p|x)=c_x\binom{n}{x}p^x(1-p)^{n-x}=c_x^*p^x(1-p)^{n-x}. \qquad (5.5)$$

此处 c_x, c_x^* 都是只与 x 有关的数. 密度(5.5)属于 Beta 分布族(见(4.39)式). 由此知 c_x^* 必为 $1/B(x+1, n+1-x)$，后验分布就是 Beta 分布 $B(x+1, n+1-x)$（参看(4.39)式前后一段话）.

例 5.2 设 $X_1, \cdots, X_n\sim N(a, \sigma^2)$，$a$ 未知而 σ 已知. 给定 a 的先验分布为：$a\sim N(\mu, \tau^2)$. 在有了样本 $x=(x_1, \cdots, x_n)$ 后，a 的后验密度为

$$h(a|x)=\frac{\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-a)^2\right]\times\frac{1}{\sqrt{2\pi}\,\tau}\exp\left(-\frac{(a-\mu)^2}{2\tau^2}\right)}{\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-a)^2\right]\times\frac{1}{\sqrt{2\pi}\,\tau}\exp\left(-\frac{(a-\mu)^2}{2\tau^2}\right)da}.$$

此式分子整理后有 $c_{x,\sigma}\exp(-(a-t)^2/2\eta^2)$，其中

$$t=\frac{\frac{n}{\sigma^2}\bar{X}+\frac{1}{\tau^2}\mu}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}},$$

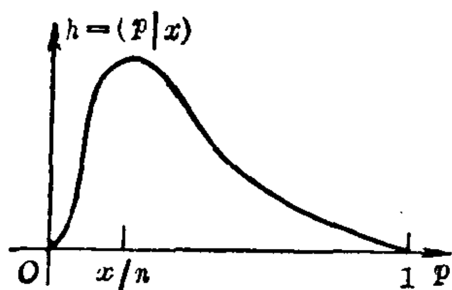
$$\eta^2=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}=\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}. \qquad (5.6)$$

因此，$h(a|x)$ 作为 a 的函数，是正态分布 $N(t, \eta^2)$ 的密度、即 a 的后验分布仍为正态分布，其均值方差由(5.6)定出.

按照 Bayes 学派的观点，后验分布综合了我们关于 θ 的先验

信息(反映在先验分布中)和样本 x 中关于 θ 的信息(这后者与样本分布有关). 如果说,先验分布概括了我们在试验前关于 θ 的认识,则经过试验得到样本 x 后,我们的认识有了变化. 现在(抽样后)我们对 θ 的认识总结在后验分布中. Bayes 学派认为,样本的唯一作用就在于它使我们对 θ 的认识起了这个转化. 这一点从(一)中那个例子看得很清楚:如果我们不知道抽出的球的颜色(即不知样本 X 的值),那我们只能说,合子为金、银、铜的可能性分别为 5/12, 4/12 和 3/12. 一旦知道了球的颜色,例如为红球,这些概率就变了:变为 70/81, 8/81, 3/81. 这时,比方说,我们对金合的可能性估计比抽球前高得多. 这些道理说起来好象很抽象,其实是很平易而常见之事. 例如,人们对某事件发展前途几个可能性有所估计(先验知识). 在获得了新的情报后,对这几种可能性大小的估计就可能改变. 比方说,原来认为不甚可能的结局,现在认为颇有可能或大有可能了.

拿例 5.1 来说,先验分布为 $p\sim R(0, 1)$,表示在抽样前,人们对 p 取各种值的可能性都认为一样大. 这个规定叫做同等无知的原则. 通俗地说,就是我们对 p 实在是无所知,只好认为它取一切值都有同等可能了[1]. 抽 n 个产品检查发现废品有 x 个,算出 p 的后验密度如上图.



在此,p 仍有可能取(0, 1)内任何值,但机会大小不处处一样了:在 x/n 这一点附近可能性最大,而接近 0, 1 处则很小. 这样我们对 p 的知识就比原来更确定了一些. 例 5.2 也可以这样解释:抽样前,对 a 的认识为 $a\sim N(\mu, \tau^2)$. 方差有 τ^2 这么大;抽样后,对 a 的认识为 $a|x\sim N(t, \eta^2)$,方差 $\eta^2<\tau^2$ (这由 (5.6) 显然). 方差愈小表示这个量散布程度小,而关于它的认识也就更确定一些[1].

1) 当然,这只是通俗的说法,从统计的观点看,p 有同等可能取一切值这一点,反映了对 p 的一种知识.

广义先验分布 按前面所说，θ 的先验分布应是一个概率分布，但有时，(5.4)式中的函数 h 非密度，但满足条件：1° $h(\theta)\geqslant 0$ 于 Θ 上. 2° 对任何样本 x，(5.4)式右边分母的积分大于 0 且有限. 这时，$h(\theta|x)$ 作为 θ 的函数，仍满足密度函数的条件($h(\theta|x)\geqslant 0$, $\int_\Theta h(\theta|x)d\theta=1$). 在这种情况下，可以形式地把 h 看成是 θ 的一个"先验密度"，而仍以(5.4)作为其后验密度. 这纯粹是一个形式的转换规则，已无条件分布的意义. 满足这种条件而非密度的 h，称为 θ 的广义先验密度. 如果有 $\int_\Theta h(\theta)d\theta=c$ 而 $0<c<\infty$，则只须以 $\frac{1}{c}h(\theta)$ 代替 $h(\theta)$，就得到真正的先验密度. 这种情况没有实质性的新东西，有意义的广义先验密度是满足

$$\int_\Theta h(\theta)d\theta=\infty \tag{5.7}$$

的那种 h. 通常把(5.7)式作为广义先验密度定义中的一个要求.

广义先验密度的作用可由下例看出：

例 5.3 设 $X_1, \cdots, X_n \sim N(\theta, 1)$，参数 θ 未知. 在抽样前，我们认为 θ 取$(-\infty, \infty)$内的一切值都是等可能的. 这可以用一个先验概率"密度"$h(\theta)\equiv 1$ 去刻划. 因为 $\int_{-\infty}^{\infty} 1\cdot d\theta=\infty$，$h\equiv 1$ 不是通常意义下的概率密度. 但若以 $h\equiv 1$, $f(x, \theta)=(2\pi)^{-n/2}\cdot \exp\left[-\sum_1^n(x_i-\theta)^2/2\right]$ 代入(5.4)式，不难算出 $h(\theta|x)=\sqrt{\frac{n}{2\pi}}\cdot \exp[-n(\bar{x}-\theta)^2/2]$，这是 $N(\bar{x}, 1/n)$的密度. 因此，$h\equiv 1$ 是一个广义先验密度.

(三) Bayes 统计推断的原则

我们已指出：对 Bayes 统计而言，除了给出样本及其分布族

1) 读者不应对这一段话产生误解，认为后验知识一定比先验知识"更确切"(比方说：θ 的先验方差一定比后验方差小). 容易举例证明并非如此. 实质之点是：后验知识是有了样本以后，关于 θ 的新认识.

外，还必须给出参数的先验分布．样本的唯一作用在于把对 θ 的认识由先验分布转化为后验分布．在这个认识的基础上，我们就可以提出下面的基本原则．

Bayes 统计推断的原则　对参数 θ 所作的任何推断（估计、检验等）必须基于且只能基于 θ 的后验分布．

此原则是这样理解的：一经由样本 x 算出了 θ 的后验分布，就设想我们除了这后验分布之外，其余的东西（样本值、样本分布、先验分布）全忘记了．这时，对 θ 作推断的唯一凭借就是这后验分布．更具体一些：如果你在对 θ 作推断的过程中需要作什么打算，或引进什么规则帮助你作推断时，这种计算只能用到后验分布，这种规则不能依赖于后验分布以外的东西，尤其是不能利用样本的分布．例如，无偏性这个规则就不能用．因为，一个估计是否无偏与样本分布族有关，不单是后验分布能决定得了的．下面的例子使我们清楚地看到这原则实质之所在．

例 5.4　矩估计不适合 Bayes 推断原则，而在一定的先验分布或广义先验分布之下，极大似然估计适合 Bayes 推断原则．

事实上，设想人家把 θ 的后验分布告诉你了，但你对样本值本身及样本分布族毫无所知．你怎样去算出样本矩和总体矩？没有这个你怎么能作矩估计？但是，若先验分布是 Θ 上的均匀分布（即 $h(\theta)\equiv 1$ 于 Θ 上，当 $\int_\Theta d\theta=\infty$ 时为广义先验分布），则不难看到：存在只与样本 x 有关的常数 $c_x>0$，使

$$h(\theta|x)=c_x f(x,\ \theta),\ \theta\in\Theta.$$

此处 $h(\theta|x)$ 为后验密度，$f(x,\theta)$ 为似然函数．MLE $\hat{\theta}$ 是使 $f(x,\theta)$（作为 θ 的函数）达到最大，这与 $h(\theta|x)$ 达到最大是一回事．既然知道了 $h(\theta|x)$，只要找 $\hat{\theta}$ 使 $h(\hat{\theta}|x)=\sup\limits_{\theta\in\Theta} h(\theta|x)$ 就行，而这个除了 $h(\theta|x)$ 外，并不需要知道其他东西．

核心之点在于（关于这一点的批评以后再谈），Bayes 主义者虽也用到样本的分布，但只是利用分布的数学形式，将其与先验分布配合以达到后验分布．古典学派对样本分布赋予一种频率解

释，因此，在进行推断时，不能忘记这样一点：手头这个样本 x 只是无限次可能的试验结果中的一具体实现. 在进行推断时，不仅要考虑到现有的 x，还要考虑到那些没有出现的可能的 x 值. 这就必须涉及样本分布. Bayes 学派反对把现有 x 放在"无限多可能值之一"这个背景下去考察. 认为：既然我们手头只有 x，我们的推断只能根据它，而不能去想到虽然可能，但我们实际没有观察到的那种 x 值. 这样，当然也就排斥了使用样本分布的可能性.

(四) Bayes 推断的具体实施

按照上段标明的基本原则，可以说，统计学家的任务就只在于算出后验分布(当然，为算出后验分布，须定出先验分布、样本分布，并抽样得到 x. 这些工作中统计学家也要起作用). 他可以把这后验分布报告使用者，让后者从中得出与他问题有关的结论. 但这样做过于笼统了. 可以从统计学的角度，针对若干常见的推断问题，提出一些供选择的、实施推断的办法.

1. 点估计　原则是：找后验分布的某个有代表性的特征数字去估计 θ. 例如，后验分布的均值或中位数. 或使后验密度 $h(\theta|x)$ 达到最大的 $\hat{\theta}$ 去估计 θ. 最后这个估计叫**广义极大似然估计**. 举几个例子.

例 5.5　考虑例 5.1. 不难算出：分布 $\mathrm{Be}(x+1, n+1-x)$ 的均值为

$$\bar{p}=\frac{x+1}{n+2}. \tag{5.8}$$

而广义极大似然估计就是通常的 MLE，即 $\hat{p}=x/n$. 暂时把关于先验分布的争论放在一边，而把(5.8)纯粹看成为对 $\hat{p}$ 的一个修正. 我们也看到，$\bar{p}$ 与 $\hat{p}$ 相比有这样的优点：当 $x=0$ 或 $x=n$ 时，$\hat{p}=0$ 或 1. 这种估计未免太极端一些，而 $\bar{p}$ 则分别为 $\frac{1}{n+2}$ 及 $\frac{n+1}{n+2}$，不为 0 也不为 1. 这个估计看上去显得合理些.

后验分布中位数要由不完全Beta分布表决定，比较复杂些.

例 5.6 考虑例 5.2. 已算出 a 的后验分布为 $N(t, \eta^2)$, t, η^2 见(5.6)式. 由正态分布的性质, 知用三种方法估计 a 的结果相同: 都是 t.

这是一个重要的 Bayes 估计. 这个估计有趣之处在于它生动地体现了这样一点: Bayes 方法综合了 θ 的先验信息与样本信息. 若只有先验信息而无样本, 则因 $\theta \sim N(\mu, \tau^2)$, 我们别无他法, 只能用 μ 估计 θ. 这是一个极端. 另一个极端是只有样本信息. 这时, 根据以前的知识, 我们知道 $\overline{X}$ 是一个良好估计. 当两种信息具备时, Bayes 估计是这两个极端的加权平均, 权的比为 $\frac{1}{\tau^2}:\frac{n}{\sigma^2}$. 这个比值也很有意思: τ^2 愈小, θ 的先验知识愈确切, 它的重要性也愈大. 故权与方差 τ^2 成反比. 又 $\overline{X}$ 的方差为 σ^2/n. 根据同样的理由, 其权应与 $\frac{n}{\sigma^2}$ 成比例. 这表明: 当先验知识很确定时, 只有极明显的现时证据才能使我们改变看法.

其他的例子我们不一一列举了, 因为这都是一些机械的计算, 没有多少兴味. 我们再提出一个多少使初学者有些困扰之点来谈一谈. 问题如下: 既然照 Bayes 学派观点看, θ 是一个有一定分布的随机变量. 那么, "估计 θ" 的确切含义是什么? 是不是说要估计一个随机变量? 如果是, "估计随机变量" 的意义何在? 回答是: *我们估计的是随机变量 θ 在一个特定场合下所取的特定值*. 这一点就废品率 p 这个例子看得很清楚: 不同的批 p 有不同的值. 在这个意义上 p 是随机的. 我们要估计的不是这个抽象的 p, 而是摆在我们面前这一批产品的 p 值. 但这一来又有下面的问题: 既如此, 则 p 是一个确定的未知值, 其后验分布作何解释? 你不妨这样想: 在抽样得到 x(n 个样品中废品个数)后, 我们并无办法确切地定出这一批的 p 值, 因而可考虑给人们这样一个回答:

"根据我们以前对 p 的了解(先验分布), 及现在观察的结果(样本 x), 我们推断: 未知的 p 有 90% 的可能性 $\leqslant 0.01$, 有 5% 的可能性在 0.01 到 0.03 之间, 有 5% 的可能性超过 0.03%".

我们想：绝大多数人都会承认，以上的陈述确实可视为问题的一种回答．即使在日常生活中这类说法也累见不鲜．如"完成这次工作，有 90% 的可能性一星期就够了．有 10% 的可能性要超过一星期"．这个说法丝毫没有否定，完成工作所需天数是一确定的未知数．这就是后验分布的含义．当然，在许多时候需要拿出一个更确定的回答．这时，后验分布均值和广义极大似然估计等提供了可能的选择．

这种论点就必然要容许概率的非频率解释．因为，从频率的观点看，说"一个确定的未知值≤1 的可能性为 0.90"，是毫无意义的．

2．假设检验　问题提法仍如(3.32)的形式，没有两样．具体检验方法如下：算出 θ 的后验分布 $P(\theta|x)$后，计算 Θ_H 和 Θ_K 的后验概率

$$p_H(x)=P(\theta\in\Theta_H|x),\ p_K(x)=P(\theta\in\Theta_K|x). \qquad (5.9)$$

其意义是：在综合先验和样本信息的情况下，发现 θ 落在 Θ_H 内的可能性为 $p_H(x)$，落在 Θ_K 内的可能性为 $p_K(x)$．因此提出这样的检验法则：

$$\text{当 } p_H(x)>1/2 \text{ 时接受 } H, \text{当 } p_H(x)<1/2 \text{ 时否定 } H. \qquad (5.10)$$

若 $p_H(x)=1/2$，则接受或否定都可以．

习惯于从 NP 理论的观点看待假设检验问题的人，可能会提出如下的异议：这个作法没有考虑到 H 和 K 的相对重要性，或更确切地说，没有考虑到两类错误后果的差异．在 NP 理论中选定水平 α 就是为了这一点．对这个异议的回答是：这反映了纯科学的"推断"与采取一定"行动"(行动有其后果)之间的差异．推断的目的是在于给问题一个看来最好的回答而不计其利害后果．拿此处的问题来说，设想你处在这样一个地位：有一个事件 A，其可能性为 0.6．在做试验前让你预测 A 发生还是不发生．若你对此预测不承担任何后果，当然会预测 A 发生．但如你这个预测是一个要担负后果的行动(例如，若预测 A 发生而实际上并未发生，损失

1000 元. 反之,若预测 A 不发生而发生了,损失 0.1 元),你的选择就不能单根据 $0.6>1/2$ 这一点了.

Bayes 方法也可以把采取行动的后果考虑进来. 这时就不一定用(5.10). 这一点到 § 5.2 再谈.

例 5.7 设样本 $X\sim N(\theta, 100)$, θ 的先验分布为 $\theta\sim N(100, 225)$. 对 X 观察一次得到 115. 要检验假设 H: $90\leqslant\theta\leqslant 110\leftrightarrow\theta<90$ 或 $\theta>110$.

按例 5.2 中的计算($n=1$, $\sigma^2=100$, $\mu=100$, $\tau^2=225$),得知当 $X=115$ 时, θ 的后验分布为 $N(110.38, 69.23)$. 于是

$$p_H(115)=\frac{1}{\sqrt{2\pi}\sqrt{69.23}}\times\int_{90}^{110}\exp[-(\theta-110.38)^2/138.46]d\theta=0.473$$

(此积分可由查正态分布表得到). 由于 $p_H(115)<1/2$, 否定原假设 H.

例 5.8 设样本 $x\sim N(\theta, 1)$, 先验分布 H 是这样的: 它在 $\theta=0$ 一点集中了概率 p_0. 剩下的概率 $1-p_0$ 按正态 $N(\mu, \tau^2)$ 分配给 $-\infty<\theta<\infty$, $\theta\neq 0$, 即在$\{-\infty<\theta<\infty, \theta\neq 0\}$这个集上有概率密度 $(1-p_0)\dfrac{1}{\sqrt{2\pi}\tau}\exp[-(\theta-\mu)^2/2\tau^2]$ ($\theta=0$ 这点是否除外不要紧,因为有密度时,一个点的概率为 0). 要检验假设 $H:\theta=0\leftrightarrow K:\theta\neq 0$.

如果先验分布的概率全部集中在 $\theta=0$ 一点,则 X 的边缘分布密度为 $\dfrac{1}{\sqrt{2\pi}}e^{-x^2/2}$; 如果先验分布的全部概率都按 $N(\mu, \tau^2)$ 的规律分配,则 X 的边缘分布密度应为

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2}\frac{1}{\sqrt{2\pi}\tau}e^{-(\theta-\mu)^2/2\tau^2}d\theta$$

$$=\frac{1}{\sqrt{2\pi(1+\tau^2)}}\exp\left[-\frac{(x-\mu)^2}{2(1+\tau^2)}\right]$$

(这是 $N(\mu, 1+\tau^2)$ 的密度). 暂记 $g(x)$ 为上式右边,而 $m(x)=p_0$

$\cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} + (1-p_0)g(x)$，则有

$$p_H(x) = P(\theta=0|x) = p_0 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Big/ m(x),$$

$$p_K(x) = P(\theta \neq 0|x) = (1-p_0)g(x)/m(x).$$

易算出

$$p_H(x) > p_K(x) \Leftrightarrow \left| x + \frac{\mu}{\tau^2} \right| < A,$$

$$A^2 = \frac{2(1+\tau^2)}{\tau^2} \left(\frac{\mu^2}{2\tau^2} + \frac{1}{2} \log(1+\tau^2) + \log \frac{p_0}{1-p_0} \right).$$

注意当 $\mu \neq 0$ 时，接受域已不与原点对称.

此例的特色在于，先验分布在一点 $\theta=0$ 处的概率大于 0，而在这点之外有密度. 因为 θ 严格地等于 0 几乎是不可能的，故 Bayes 学派认为，只有赋予 0 这个点以大于 0 的先验概率，问题才是有意义的.

3. 区间估计　在求得 θ 的后验分布 $P(\theta|x)$ 后，找区间 $[A(x), B(x)]$，使

$$P(A(x) \leqslant \theta \leqslant B(x) | x) = 1-\alpha. \tag{5.11}$$

$\alpha \in (0, 1)$ 为给定的数. (5.11)式中的 $1-\alpha$ 在 Bayes 统计中没有专名，故称之为**后验置信度**. 适合(5.11)的任何 $[A(x), B(x)]$，称为 θ 的后验置信度 $1-\alpha$ 的区间估计. 一般，适合(5.11)的区间 $[A(x), B(x)]$ 很多，可以挑选其中长度最短者.

例 5.9　设 $X_1, \cdots, X_n \sim N(\theta, \sigma^2)$，$\sigma^2>0$ 已知，θ 为参数. θ 的先验分布为 $N(\mu, \tau^2)$. 据例 5.2，在已有样本 $X_1, \cdots, X_n$ 时，θ 的后验分布为 $N(t, \eta^2)$，t 和 η^2 由(5.6)式给出. 由此易知，后验置信度为 $1-\alpha$ 的区间估计是 $[t-\eta u_{\alpha/2}, t+\eta u_{\alpha/2}]$.

例 5.10　设 $X \sim B(n, p)$，且 p 的先验分布为 $R(0, 1)$. 在例 5.1 中已得到 p 的后验密度为(5.5)式. 此密度作为 p 的函数先升后降，由此易知，后验置信度为 $1-\alpha$ 的区间估计 $[p_1(x), p_2(x)]$ 由以下的关系式确定(简记 $p_i = p_i(x)$, $i=1, 2$):

$$p_1^x(1-p_1)^{n-x} = p_2^x(1-p_2)^{n-x},$$

$$\int_{p_1}^{p_2} p^x(1-p)^{n-x}dp=(1-\alpha)B(x+1,\ n+1-x).$$

例 5.11 设 $X_1, \cdots, X_n$ 为从具 Cauchy 分布密度

$$f(x,\ \theta)=\frac{1}{\pi(1+(x-\theta)^2)},\quad -\infty<x<\infty,\ -\infty<\theta<\infty \tag{5.12}$$

中抽出的独立随机样本，θ 的先验分布为 $N(0,\ 1)$. 则 θ 的后验密度为(记 $x=(x_1, \cdots, x_n)$)

$$h(\theta|x)=c_x e^{-\theta^2/2}\prod_{i=1}^{n}[1+(x_i-\theta)^2]^{-1},$$

$$c_x=\left(\int_{-\infty}^{\infty} e^{-\theta^2/2}\prod_{i=1}^{n}[1+(x_i-\theta)^2]^{-1}d\theta\right)^{-1}.$$

函数 $h(\theta|x)$(变元为 θ)不是单峰的，因此找最短区间估计的问题要比前几例复杂. 这例只能用数值解法. c_x 的计算需用数值积分. 然后，可以考虑取以样本中位数 m_n 为中心的区间 $m_n\pm d$. 当然，这一来就不能严格遵守区间最短的要求.

以上我们介绍了几种基本类型的统计推断问题的 Bayes 处理法，其他问题也是按照这个格式处理. 比方说，θ 的参数空间 Θ 分解为 $k\geqslant 2$ 个两两不相交的部分 $\Theta_1, \cdots, \Theta_k$，统计推断问题是要在 k 个命题 $H_i: \theta\in\Theta_i,\ i=1, \cdots, k$ 中选择其一. 有了样本 x 后，根据样本分布和先验分布算出 θ 的后验分布 $P(\theta|x)$，由此算出命题 H_i 的后验概率 $p_i(x)=P(\theta\in\Theta_i|x),\ i=1, \cdots, k$，如果

$$p_{i_0}(x)=\max_{1\leqslant i\leqslant k} p_i(x),$$

则选择 H_{i_0}，即认为 $\theta\in\Theta_{i_0}$. 自然，实际问题一般不象前面的例子那么单纯. 而往往有一些更细致的问题要考虑，甚至不时有必要对 Bayes 方法的含义作某种引伸. 这些都不能在此细述了. 有兴趣的读者可参看 Box 和刁锦寰合著的《Bayesian Inference In Statistical Analysis》.

(五) 确定先验分布的方法

用 Bayes 方法处理统计推断问题的前提是给出 θ 的先验分

布. 因此,怎样去确定先验分布的问题是一个至关重要的问题.在文献中提到过的方法,主要有以下一些.

1. 客观法 在象前面提到的废品率 p 的那类例子中,参数本身确有一种可赋予频率解释的随机性, 且如对以往的资料有些积累,则可以由之对先验分布作出较准确的估计. 在这种情况下,先验分布可通过这种估计得到. 由于这种方法看来没有渗杂多少人的主观因素,姑且把它称之为客观法.

一般认为,即使某个统计学者对 Bayes 统计持否定态度,他也不反对在先验分布可以"客观"地定出的条件下, 使用如(四)中介绍的推断方法去处理统计问题. 就"纯正"的 Bayes 学派而言, 他们也不反对在上述情况下确定先验分布的方法, 但他们反对把这种方法赋予频率解释, 更不赞成在褒扬的意义下使用"客观"一词.

在不少情况下, 以往积累的资料并不是直接给出了参数在当时的值,而只是其一种估计. 如在废品率的例中,以往的资料不必逐日给出废品率 p 的值. 更可能的是: 逐日进行了抽样. 每日抽 N 个,记录了其中的废品数 X_1, X_2, …等. 利用它们也可以对先验分布进行估计,或更一般地, 这种资料可以以某种方式用于 Bayes 推断. 这种方法叫做经验 Bayes 方法, 最早由 H. Robbins 在 1955 年提出来.

2. 主观概率法 按 Bayes 学派的说法, 这是一种通过"自我反省"去确定先验分布的方法. 就是说,对参数 θ 取某某值的可能性多大, 通过思考, 觉得该如何, 而定下一个值. 有人用一种"打睹"的设想作了形象的解释. 设 θ 可能值的区间为 $0\leqslant\theta\leqslant1$, 先一分为二: A: $0\leqslant\theta\leqslant1/2$, B: $\frac{1}{2}<\theta\leqslant1$. 有人要和你打睹是 A 出现还是 B 出现. 如果你经过"反省"定下这样一个数 a, 使得对任何 $b\leqslant a$ 你都愿意以 $1:b$ 的输赢与他打睹(即: 若 A 出现你得 1 元,否则付出 b 元),则表示你认为 A 的可能性, 即 $0\leqslant\theta\leqslant1/2$ 的概率是 $\frac{a}{1+a}$.

本书作者觉得，主观先验分布应当是反映了个人以往对 θ 的了解，包括经验知识和理论知识. 这里面自然也有由其他人获得而经他吸收的经验和理论知识. 这种知识可能没有经过适当的组织和整理. 当他需要对 θ 给出先验分布时，他经过"自我反省"将这些经验和知识作了整理. 这样提出的先验分布，就有可能较确切地反映了他目前关于 θ 的知识. 这当然还不能保证，所提出的先验分布按某种客观标准(如果可以提出这种标准的话)是正确的. 只有这样去理解主观概率法才有意义. 至于上面提到的"打赌"一说，作者认为是没有什么意义的. 充其量不过是一种循环式的论证而已，并不能提供什么新东西.

3. *同等无知原则* 这原则有时称为 Bayes 假定. 以废品率 p 为例. 当我们对 p 并无所知时，我们只好先验地认为，p 以同等机会取(0, 1)内各种值，因而以(0, 1)内均匀分布 $R(0, 1)$ 作为 p 的先验分布. 这就是所谓同等无知原则.

这个原则有一个困难. 就拿上例来说，如果我们对 p 无所知，则对 p^3 也无所知. 因此按同等无知的原则，也可以取 $R(0, 1)$ 为 p^3 的先验分布. 但这时 p 的先验分布就不是 $R(0, 1)$ 了.

4. *无信息先验分布* 这个方法在某种意义上类似于同等无知原则. 但由于它只能用于某些特殊类型的分布族的参数，因而就没有上面提到的那种困难. 我们举几个例子来说明这个方法.

例如，设总体分布有密度函数 $f(x-\theta)$，$-\infty<\theta<\infty$. θ 为一个位置参数. 若将度量原点由 0 移至 $-c$，则总体变量的密度变为 $f(x-(\theta+c))$. 如果先验分布不依赖于原点的选择(这就是本例中"无信息"一词的含义)，则它在等长区间内的先验概率应当一样，换句话说，先验密度应当恒等于 1. 这就是本例中的无信息先验分布，它是一个广义先验分布.

又如，设总体分布有密度函数 $\frac{1}{\theta}f\left(\frac{x}{\theta}\right)$，$\theta>0$. 这种参数 θ 称为刻度参数. 因为，若将度量单位由 1 改为 $\frac{1}{c}$，则总体变量的密度变为 $\frac{1}{c\theta}f\left(\frac{x}{c\theta}\right)$. 如果先验分布不依赖于刻度的选择(这是

本例中“无信息”的含义），则对任何 a, b, c, $0<a<b$, $c>0$, θ 落在 $[a, b]$ 内的先验概率，应等于其落在 $[ca, cb]$ 内的先验概率. 不难看出，这只有在先验密度为 $\frac{1}{\theta}$ (当 $\theta>0$. $\theta<0$ 时为 0)时才可能. 这就是此例中的无信息先验分布，它也是一个广义先验分布.

由以上两例可以窥见此法的思想：设总体变量 X 的分布族在某种变换下不变（以上两例中，变换分别是 $X\to X+c(-\infty<c<\infty)$ 和 $X\to cX(0<c<\infty)$），则先验分布取成有相应的不变性. 不可拘泥于“无信息”的字面意义（这很难说清楚），而不如把它理解为一种与分布族的特定结构（反映在对一定的变换保持不变）相适应的先验分布. 直观上人们觉得，这是一个较稳妥的选择，而可能带来较好的结果.

5. 共轭先验分布　这是一个基于纯数学考虑的选择原则. 定义如下：设 $\mathscr{F}$ 为 θ 的一个先验分布族. 如果对任取的 $H\in\mathscr{F}$ 及样本值 x，后验分布总属于 $\mathscr{F}$，则称 $\mathscr{F}$ 是一个共轭先验分布族. 由于后验分布不仅依赖于 H 和 x，还依赖于样本分布族. 因此，某一指定的先验分布族是否有共轭性，要视样本分布族而定.

如在例 5.2 中，$X_1, \cdots, X_n \sim N(a, \sigma^2)$. 当 a 的先验分布为正态分布时，不论样本值如何，a 的后验分布总是正态的. 故在此例中，正态分布族是一个共轭先验分布族. 又如例 5.1. 若取 p 的先验分布为 Beta 分布 $Be(a, b)$, $a>0$, $b>0$, 则易见 p 的后验分布为 $Be(x+a, n+b-x)$. 这说明，Beta 分布族 $\{Be(a, b): a>0, b>0\}$ 是一个共轭先验分布族. 下面再举几个例子. 这些例子说明了在充分统计量存在的情况下，确定共轭先验分布族的一般方法.

例 5.12　设 $X_1, \cdots, X_n \sim N(a, \sigma^2)$, a 已知而 σ 为未知参数. 求 σ 的共轭先验分布族.

第一步是写出样本的概率密度 $(2\pi)^{-n/2}\sigma^{-n}\exp\left(-\sum_{i=1}^{n}(x_i-a)^2/2\sigma^2\right)$. $\sum_{i=1}^{n}(x_i-a)^2$ 为充分统计量，记之为 T（注意 $T>0$）. 然后，

确定那些 d, $d>0$, 使

$$0<\int_0^\infty \sigma^{-d}\exp(-T/2\sigma^2)d\sigma<\infty.$$

显然，一切 $d>1$ 都适合这个要求，且不难算出上述积分值为(作变数代换 $x=T/2\sigma^2$)$2\left(\frac{T}{2}\right)^{(d-1)/2}\Big/\Gamma\left(\frac{a-1}{2}\right)$. 以 $D(d, T)$ 记一分布，其密度为

$$\left[\Gamma\left(\frac{d-1}{2}\right)\Big/2b^{(d-1)/2}\right]\sigma^{-d}\exp(-b/\sigma^2).$$

$$(\text{当 }\sigma>0.\ \sigma<0\text{ 时为 }0) \tag{5.13}$$

则 $\{D(d, b): d>1, b>0\}$ 为 σ 的一个共轭先验分布族. 事实上很容易看出，若 σ 有先验分布 $D(d, b)$，样本 $x=(x_1, \cdots, x_n)$. 则 σ 的后验分布为 $D(d+n, b+T/2)$.

例 5.13 设 $X_1, \cdots, X_n\sim N(a, \sigma^2)$, a 和 σ 都未知. 要确定 (a, σ) 的共轭先验分布族.

第一步仍是写出样本的概率密度如上例. 这密度函数可写为 $(2\pi)^{-n/2}\sigma^{-n}e^{-T/2\sigma^2}e^{-n(\bar{x}-a)/2\sigma^2}$. 如果把此函数乘以一个适当的常数，则可得到 (a, σ) 的一个密度函数. 细察此密度函数，知其有如下的特点: 1° σ 的边缘分布为 $D(n, T/2)$. 2° 在给定 σ 的条件下，a 的条件分布为 $N(a, \sigma^2/n)$. 以 $G(d, b, \mu, \tau)$ 记 (a, σ) 的这样一个分布，其中 σ 的边缘分布为 $D(d, b)$，而在给定 σ 时，a 的条件分布为 $N(\mu, \sigma^2/\tau)$. 记 $\mathscr{F}=\{G(d, b, \mu, \tau): d>1, b>0, -\infty<\mu<\infty, \tau>0\}$，则易见 $\mathscr{F}$ 为 (a, σ) 的一个共轭先验分布族. 事实上，若 (a, σ) 的先验分布为 $G(d, b, \mu, \tau)$，则 $(a, \sigma, X_1, \cdots, X_n)$ 的联合密度为

$$C\sigma^{-d}\exp(-b/\sigma^2)\sigma^{-1}\exp(-\tau(a-\mu)^2/2\sigma^2)\sigma^{-n}$$
$$\cdot\exp\left(-\sum_{i=1}^n(x_i-a)^2/2\sigma^2\right). \tag{5.14}$$

C 为一个与 a 和 σ 都无关的常数. 因为

$$\tau(a-\mu)^2+\sum_{i=1}^n(x_i-a)^2=(n+\tau)(a-t)^2$$

$$+n\tau(\bar{x}-\mu)^2/(n+\tau)+T.$$

其中 $t=(n\bar{x}+\tau\mu)/(n+\tau)$，知(5.14)式可写为

$$C_1\sigma^{-(d+n-1)}\exp\left[-\left(b+\frac{T}{2}+\frac{n\tau(\bar{x}-\mu)^2}{n+\tau}\right)\Big/\sigma^2\right]$$

$$\cdot\frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n+\tau}}\exp\left(-\frac{(a-t)^2}{2\sigma^2/(n+\tau)}\right).$$

C_1 为一个与 a 和 σ 都无关的常数．细察此式，知它就是 $G\left(d+n-1,\ b+\frac{T}{2},\ t,\ \frac{\sigma^2}{n+\tau}\right)$，仍属于 $\mathscr{F}$．这证明了 $\mathscr{F}$ 是共轭先验分布族．

例 5.14 设总体有 Poisson 分布：$P_\theta(x)=e^{-\theta}\theta^x/x!$，$x=0, 1, 2, \cdots$，$\theta>0$ 为参数．设 $X_1, \cdots, X_n$ 为抽自此总体的独立随机样本，则其概率函数为

$$\prod_{i=1}^{n}(e^{-\theta}\theta^{x_i}/x_i!)=e^{-n\theta}\theta^S/(x_1!\cdots x_n!),\ S=\sum_{i=1}^{n}x_i. \qquad (5.15)$$

以 $G(a,b)$ 记带参数 a, b 的 Gamma 分布，即有密度函数 $(a^b/\Gamma(b))x^{b-1}e^{-ax}$（当 $x>0$，$x<0$ 时为 0），则由(5.15)的形式易见：Gamma 分布族 $\mathscr{F}=\{G(a, b): a>0, b>0\}$ 是 σ 的一个共轭先验分布族．事实上很容易验证，若 θ 有先验分布 $G(a, b)$ 而样本值为 $x_1, \cdots, x_n$，则 θ 的后验分布为 $G(a+n, b+S)$．

共轭先验分布并不是从 θ 取种种值的可能性大小的考虑出发，而纯粹是为了数学上的方便．如我们在 § 5.2 中将看到的，有时我们赋予参数以一定的先验分布，倒不是想要用 Bayes 统计的观点去处理有关的统计推断问题，而是为了在数学上证明某种结果，或作为求具有某种优良性质的解的手段．这时，任何先验分布(不管它是否符合实际)，只要有助于达到这个目的的，都可以用．共轭先验分布由于其特有性质，常是合宜的候选者．另外，在有些场合下，种种选定先验分布的法则都不能给出适当的结果，这时，共轭先验分布由于其在计算上的简单方便，而被人们所选用．

（六） Bayes 学派与频率(古典)学派的争论

频率学派和 Bayes 学派是当今数理统计学的两大学派. 统而言之,凡是坚持概率的频率解释,因而对数理统计学中的概念、结果、方法性能的评价等,都必须在大量重复的意义上去理解的,都属于频率学派. 自本世纪初数理统计大发展以来,一些起领导作用的学者,如 Fisher, K. Pearson, J.Neyman, E. S. Pearson 等,都属于这个学派[1]. 因此直到大约五十年代为止,这个学派占据了主导的地位. 近二、三十年以来,Bayes 学派迅速崛起,达到了可与频率学派分庭抗礼的程度. 与频率学派相比,其发展较新,因此人们,特别是 Bayes 学派中人,往往也把频率学派称为古典学派.

这两个学派之间经常而热烈的争论,是当代数理统计学发展的一个特有的现象. 为此而发表了许多文章和言论. 双方的主要论点已由学者们充分阐明过了,但至今并无定论. 不过,似乎双方都不否定. 这两个学派的方法在许多具体问题中的应用,都给出了一些有益的结果. 尽管对此事的解释各有不同,在 Bayes 学派中有些人看来,频率学派中的一些重要方法之所以能站住脚,只是因为它暗合于某个合理的 Bayes 解. 例如,$N(\theta,1)$中 θ 的无偏估计 $\bar{X}$,恰好是当 θ 有(广义)先验密度 $h(\theta)\equiv1$ 时的 Bayes 估计. 在本例中,$h(\theta)\equiv1$ 是所谓"无信息先验分布",因而 $\bar{X}$ 是一个合理的 Bayes 解. 也有人认为: $N(a, \sigma^2)$中参数 a 的 t 区间估计之所以能被接受,也只是因为它是一定的(广义)先验分布下的 Bayes 解. 频率学派也承认 Bayes 方法在一些情况下可用,但限于先验分布可给予某种"客观"解释(实际上就是频率解释)的时候. 至于两派涉及的基本哲学观点,看来是无法调和的. 很可能

---

1) Fisher 的思想比较复杂,大体上说,可以把他归入频率学派. 但他的"Fiducial inference" 明显地背离了频率学派的基本原则. 他对 Bayes 学派总的是持否定态度的,但他的某些言论却有着 Bayes 学派的色彩. 要仔细了解这个问题,有必要阅读 Fisher 的原著,尤其是他与 Jeffreys 等人争论的文章.

在应用中的表现以及广大的使用统计方法的人的倾向性，将是决定这两个学派的争论的前途的决定性因素.

频率学派对 Bayes 学派的批评，主要集中在所谓主观概率以及与之相关的先验分布的确定问题上. 按频率学派，一个事件的概率可以用大量重复试验之下事件的频率来解释，这种解释不取决于认识主体. 而主观概率则理解为认识主体对事件发生的相信程度. 坚持频率解释的人认为这不仅难以捉摸，且与认识主体有关，没有客观性，因而也就没有科学性. 以此，凡是不能给以客观的频率解释的那种先验分布，都是主观随意性的产物，是不可接受的. 当然也不能接受那种建立在这个基础上的统计方法，认为这样作出的推断缺乏客观的科学价值. 例如，Fisher 在提出"信任分布"的概念时，就特别将它与 Bayes 统计中的"后验分布"划清界线，认为前者不须对参数作任何先验的假定，而后者则必须作这样的假定.

Bayes 学派对这种批评的回答，归纳起来有这样几点.

1. 主观概率事实上是人们常用的一个概念. 例如，人们常说"明天下雨的可能性是 2/3"这类的话. 这话没有频率解释，但普遍觉得它有一种可理解的意义. 它反映了说话者对"明天下雨"这件事的相信程度. 甚至在科学上也有这种说法，例如，按照目前积累的探测结果，人们认为"火星上有生命"的可能性很低，也许不到万分之一. 因此，赞成主观概率概念的人，常把它说成是反映了说话者对有关事件的知识水平(即根据他现有的知识而作出的判断). 这话看来有一定道理，不过仍未能完全解决这个问题：主观概率的实质是什么，能否给予严格的定义，它是否真能起任何有用的作用，等等.

2. 在涉及到采取行动且必须为此承担后果的问题(所谓统计判决问题，见 § 5.2)中，人们了解的情况不同，对问题所具有的知识不同(这会反映到所采用的先验分布不同)，他们的最佳行动方案也应有所不同. 在这种情况下，不同的人有不同的先验分布是正常的，要求所谓"客观性"反倒没有意义.

3. Bayes 学派认为，虽则古典统计没有明白使用先验分布，但事实上，在频率学派观点之下导出的统计推断方法，也是某种潜在的先验分布之下的 Bayes 解. 前面我们已提到过两个这样的例子. 在不少情况下，这个事实上的先验分布往往很不合理. 例 5.21 就是一个例子. 因此，Bayes 学派认为，与其不顾这个事实而否定先验分布，不如明确承认它的存在，反而有可能选用较好的先验分布.

从(四)段中的一些例子看到，Bayes 方法有一个优点，就是它求解的程序简单：只要算出了后验分布，则解就可以得出. 纵使有什么困难，也是计算性质的. 不象在古典统计中，问题的解决往往取决于推导出复杂的抽样分布. 这个特点受到理论训练较少的应用者的欢迎. 有的频率学派的学者对此有所批评，认为这是把问题人为地简单化了且容易导至滥用. Bayes 学派则认为，Bayes 方法抓着了本质的东西而避免了那些无谓的复杂细节(Bayes 学派既然否定频率观点，自然就不认为抽样分布有什么意义). 至于它可能被滥用，那是使用者的问题，与方法本身无关. 何况，古典统计方法也存在被滥用的问题.

频率学派对 Bayes 学派还有这样一个批评：Bayes 方法也要以样本分布为出发点，这种分布通常都是在频率的意义上来解释的. 因此，Bayes 学派既彻底否定频率学派，但又要使用这个学派的工具. 对于这个批评，Bayes 学派很少作出回答. 可能是因为，这个不一致性确是一个难于作出信服回答的问题. 如果作一个彻底的主观概率论者，就必须把样本分布看成是刻划样本取各种值在主观上的相信程度. 即使这样也还不能解决问题，因为样本是已知的，而 Bayes 学派反对把已有样本放到无穷多可能样本的背景下去考察这种做法(这将导致频率解释). 故推到极端，人们甚至不能谈论样本有什么分布的问题. 因此，Bayes 学派只能把样本分布作为其方法结构中的一个组成部分，而避免去涉及对它的意义的解释问题.

现在反过来谈谈 Bayes 学派对频率学派的一些批评

批评之一 涉及"频率解释"本身. 许多应用问题是一次性的. 在严格相同甚至大致相同的条件下的重复, 事实上不可能. 因此, 在许多情况下, 统计概念和方法的频率解释完全没有现实意义. 这种频率解释的根源, 来自于把样本放在"无穷多可能值"的背景下去考察这一点. Bayes 学派认为, 只能在现有样本的基础上去处理问题, 而不能去顾及那种可能发生但其实并未发生的情况. 这个批评在许多情况下是中肯的. 问题在于: 用"相信程度"去取代频率解释是否真的就克服了这个困难.

批评之二 与上一点密切相关. Bayes 学派认为, 由于古典统计基于概率的频率解释. 因此, 所导出的方法(点估计、区间估计、假设检验等)的精度和可靠度也只是大量重复下的平均值. 这是名义性的, 且是在事前(抽样前)就已定下了的, 故称为**事前精度(可靠度)**. Bayes 学派认为, 这种不顾实际的样本值而在事前规定的精度和不靠度是不合理的, 且往往与实际情况大相径庭. 直观上人们倾向于能接受的是: 统计推断的精度或可靠度如何, 应与试验结果(样本)有关, 即应当采用**事后精度**和**可靠度**. Bayes 学派形式上符合这个要求.

例如, 为检验一个假设 $\theta\in\Theta_H\leftrightarrow\theta\in\Theta_K$, 选定水平 $\alpha=0.05$. 若原假设被否定了, 则我们只知道犯错误的机会是 0.05, 与所得样本无关. 而事实上人们觉得, 若问题是检验 $N(\theta,1)$ 中的 $\theta\leqslant 0$, 则在 $X=2$ 时否定原假设, 与在 $X=4$ 时否定原假设相比, 后者出错的危险要小些. 在 Bayes 方法, 则我们算出 $P(\theta\leqslant 0|X)$. 当 $X=4$ 时, 此条件概率远小于其当 $X=2$ 时之值. 因此虽则二者的结果都是否定 $\theta\leqslant 0$, 其可靠程度则有异. 在区间估计以及点估计方面也有类似问题. 下面的例子给人深刻的印象.

例 5.15 设 $X_1, \cdots, X_n \sim R\left(\theta-\frac{1}{2},\theta+\frac{1}{2}\right)$, $-\infty<\theta<\infty$. 要作 θ 的区间估计. 取置信系数 $1-\alpha=0.95$, 令 $\hat{\theta}_n=(\max\limits_{1\leqslant i\leqslant n} X_i+\min\limits_{1\leqslant i\leqslant n} X_i)/2$. 由于 $\hat{\theta}_n-\theta$ 的分布与 θ 无关, 可找到常数 c_n, 使

$P_\theta(|\hat{\theta}_n-\theta|\leqslant c_n)=0.95$，于是得 θ 的置信系数 0.95 的置信区间为 $[\hat{\theta}_n-c_n,\ \hat{\theta}_n+c_n]$. 这区间的可靠度 0.95，以及反映精度的区间长 $2c_n$，都是前定的. 对 $n=25$ 有 $c_{25}=0.056$，区间长为 0.112.

现设有两组大小各为 25 的样本:

$$甲: \max_{1\leqslant i\leqslant 25} X_i=3.278,\ \min_{1\leqslant i\leqslant 25} X_i=3.275;$$

$$乙: \max_{1\leqslant i\leqslant 25} X_i=1.256,\ \min_{1\leqslant i\leqslant 25} X_i=0.261.$$

对样本甲而言，在它的基础上，未知参数 θ 仍可以在区间 $[3.278-0.5,\ 3.275+0.5]=[2.778,\ 3.775]$ 内活动. 一个长为 0.112 的区间，只是这区间长度的 $100\ \dfrac{0.112}{0.997}\ \%=0.112$. 因此，如某人获得样本甲，那么，按 Neyman 理论算出的，名义上可靠度为 0.95 的区间，事实上可靠度不过 0.11 而已.

反之，如某人获得样本乙，则在它的基础上，我们可以百分之百地肯定未知参数 θ 不会超出区间 $[1.256-0.5,\ 0.261+0.5]=[0.756,\ 0.761]$，其长只有 0.03. 而按 Neyman 理论算出的区间为 $[0.703,\ 0.815]$. 这个区间包含了 $[0.756,\ 0.761]$，其长几为后者的四倍，而可靠度反而只有 0.95. 这显然是不合理的.

若用 Bayes 方法就没有这个问题. 如在样本乙的情况，则对任何先验分布而言，后验分布的概率将全集中在区间 $[0.756,\ 0.761]$ 内. 因此，这个区间的后验置信度为 1. 若在样本甲，且规定对很大的 $a>0$，θ 的先验分布为均匀分布 $R(0,\ a)$，则后验分布将是 $[2.778,\ 3.775]$ 内的均匀分布，一个后验置信度为 0.95 的区间，其长将达到 0.947，而不是由 Neyman 理论给出的 0.112. 因此，Bayes 方法所给的事后精度和可靠度符合实际的看法. 我们留给读者去讨论：在本例中 θ 的点估计 $\hat{\theta}_n$ 的精度也有类似现象.

Bayes 学派关于事前精度(可靠度)的批评，确是揭示了频率学派的一个重要缺陷. 在有些情况下，这个缺陷可以在古典统计的范围内得到补救，例如拟合优度检验. 如果不止是给出一个"接受"或"否定"的回答，而是给出拟合优度 $P(k\geqslant k_0)$ (见 § 3.2)，则无论是接受或否定原假设，这个数可以提供所作结论的可靠性. 对

于 $N(\theta, 1)$ 中检验 $\theta\leqslant 0$ 的问题显然也可作类似处理：算出

$$\max_{\theta\leqslant 0} P_\theta(X>2)=P_0(X>2) \text{ 和 } \max_{\theta\leqslant 0} P_\theta(X>4)=P_0(X>4).$$

后者比前者小很多，因而在 $X=4$ 时，否定 $\theta\leqslant 0$ 的结论，要比在 $X=2$ 时作出同一结论可靠得多．对区间估计和点估计的情况则没有类似的简单补救办法．

以上所介绍的就是这两个学派争论的一些主要问题．还有一些其他的问题，在此不一一细述了．对这种问题，目前意见还比较分歧，读者应以批判的眼光对待种种论点，以形成自己的看法．

§5.2 统计判决理论

（一）引言

统计学家 A. Wald 在 1950 年出版了一本题为«Statistical Decision Function»的著作，建立了一种统一地处理各种统计问题的理论，世称统计判决(或决策)理论．一般把 Wald 上述著作出版之日确定为这理论诞生之时，但在 Neyman 和 Pearson 于二、三十年代所发展的假设检验理论中，已包含这个理论的某些概念的萌芽，特别是把统计问题提成一个数学最优化问题的解这一点，或许对 Wald 建立其理论有所启发．

这个理论的建立是战后数理统计学发展的重大事件．从理论上说，它通过将统计问题提成数学最优化问题的解，并引进形形色色的优良性准则，不仅加速了数理统计学的纯数学化趋势的一面(这种趋势是否有益和可取是一有争议的问题)，且开拓了某些传统分支的研究领域和提出了一些有意义的统计问题．例如，参数估计这个分支在判决理论的影响下，战后的面貌有了很大的变化．其他统计分支也程度不同地受到这个理论的影响．从应用上说，这个理论通过引进“行动”的概念(用一个数值或区间去估计一个未知参数，对一个假设检验问题采取“接受”或“否定”的决定，都可看作是一种行动)，以及因行动而承担后果——通过一个被称作损失函数的数量去表达，把数理统计学的任务，推出了姑且可称之为

“发现事实”或“探求科学真理”的单纯推断的范围．而达到以追求更大的经济利益为目标的决策的领域，使数理统计学具有更大的应用意义．

本节的目的是对 Wald 理论的基本要点作一个简单的介绍．这个理论可以说不属于我们所设想的统计基础课的范围．但为了对数理统计基础有一个较全面的认识，了解这理论的一些初步知识是有必要的．

（二）统计判决问题的三要素

设想一个经济领导工作者某甲要对其业务有关的问题作出一项决策．决策的内容可以是关于某一种产品应生产多少，或是某一产品价格应定为多少以至是否将该产品投放市场等．这种决策是根据种种考虑作出，也不一定就要涉及统计方法．如果它适合以下的条件，那就必然与统计方法有关，而可以称之为统计决策问题．条件是：在作出决策所依据的事实中，有一个组成部分是受到随机性影响的观察或试验数据 X（样本）．例如，在决定某一种产品的生产规模时，过去一段时间内该产品的销售记录是依据之一．而这种记录是受到随机性影响的．样本 X 有概率分布 $F_\theta(x)$，依赖于某个参数 θ，θ 取值于参数空间 Θ．为了易于理解，我们还不妨假定，一旦知道了 X 的分布，即参数 θ 的值，则能够明确应该采取的决定．

这样，某甲在作出其决定时，必须有样本 X 且规定了样本 X 的概率分布族 $\{F_\theta, \theta\in\Theta\}$．这构成统计判决问题的第一个要素．

其次，某甲还必须在事前就明确：他能作的决定有那一些．如在上例中，若在考虑的时期内该产品最大的生产能力为 a，则某甲可以决定任一数 $d\in[0, a]$，而把产量定为 d．这时，他能作的决定的全体，可以用区间 $[0, a]$ 来表达．在统计判决理论中，常把一个可能的决定（如上文的 d）称为一个行动．而一切可能的决定的全体，如上文的 $[a, b]$，则称为该问题的行动空间，通常用 $\mathscr{D}$ 来记．又如，某甲面临的决策是：是否投放该产品．他能采取的决定有

二：投放，或不投放．这分别可用代号 0 和 1 来记．这时，问题的行动空间为 $\mathscr{D}=\{0, 1\}$．

行动空间是统计判决问题的第二个要素．这个要素确定了所要解决的问题的性质．因为，只要把全部可能的行动都罗列出来，我们就会明白要解决的是什么问题．

最后，某甲还必须明确，他在采取这样或那样的行动时，在种种情况下可能产生的后果．这后果不论其性质如何，都必须最终地归结为一个非负数量——比方说，经济上损失的人民币的元数．例如，若某甲决定的生产数量过大而市场滞销，则将造成积压．需要明确每积压一件损失多少元．如生产数量过少，则利润有损失，要确定每少销一件损失多少元．又如，他要作的决定是，是否把一批可能有些问题的产品投放市场．若有问题而投放，则信誉上有所损失．这种损失也必须经由一定的考虑转换为经济上的损失．

我们规定：若是知道了样本 X 的分布参数 θ 的值，则行动空间 $\mathscr{D}$ 中每一个行动 d 所导致的损失也就确定了．换句话说，我们规定：损失是参数 θ 和行动 d 的函数，记为 $L(d, \theta)$．这个函数称为**损失函数**，是统计判决问题的第三个要素．

例 5.16 某商店每日从工厂进货一批计 N 件．商店从该批中抽取 n 件检验，根据这 n 件中的废品数 X，决定是否接受该批产品．若接受(行动 d_1)，则每件废品商店损失 10 元．若不接受(行动 d_2)，则该店当日无货出售，每件损失利润 2 元．要决定其损失函数．

假定检验为非破坏性的，且 $n/N\approx 0$，因而可以认为 $X\sim B(n, p)$，参数空间为 $0\leqslant p\leqslant 1$．这就是本问题的样本分布族．行动空间 $\mathscr{D}=\{d_1, d_2\}$．若已知废品率 p，则该批中有废品 Np 件，商店接受这些废品造成损失 $10Np$．若不接受，则放弃了该批中的 $N(1-p)$ 件合格品，损失利润 $2N(1-p)$．由此得到本问题的损失函数为

$$L(d_i, p)=\begin{cases}10Np, & \text{当 } i=1,\\ 2N(1-p), & \text{当 } i=2,\end{cases}\quad 0\leqslant p\leqslant 1. \tag{5.16}$$

在许多实际问题中损失的计算可能不象这么简单明了. 事实上，怎样去确定损失函数使之与实际情况符合，在应用上是一件很重要且困难之事，需要做必要的调查研究工作. 在此我们强调的是：损失必须能由参数值 θ 及行动 d 所确定. 若经过研究发现损失事实上与 θ 无关或关系很小，那就说明：所观察的对象（样本 X）事实上与要解决的问题无甚关系. 这种观察结果就对问题无用了. 完全可能：行动 d 造成的损失不仅与 θ 有关，还与其他一些因素有关. 如果是这样，我们就必须假定，这些其他因素已通过某种考虑固定下来，因而行动 d 的损失只取决于 θ. 若这些因素无法固定，则必须设法取得其观察值，把 θ 的含义扩大，以达到使损失只与 θ 有关的目的.

在论述了统计判决问题的三要素后，我们就自然而然地会想到：统计判决问题的解决，就是要根据样本 X 选择一定的行动 d，使损失尽可能地小. 因为，若知道了 θ，就可以由损失函数 $L(d, \theta)$ 的形式看出何种行动最好，而样本 X 中包含了关于参数 θ 的信息，当然有助于 d 的选择.

统计判决问题的这种提法，使统计与博弈论发生了联系. 可以把大自然作为博弈的一方，她掌握了 θ 的秘密；统计学家为另一方，他力求所采取的行动 d 能获致尽可能大的利益（利益是损失的反面。因此，也可以用“利益函数”代替损失函数，问题由“使损失最小”改为“使利益最大”）. 不同之处在于：大自然不是一个自觉的行动者，不能说大自然的目标是使统计学家的损失尽可能大.

还有一点要提到一下：此处没有把 θ 的先验分布作为统计判决问题的要素之一，因为人们可以也可以不采取 Bayes 学派的观点来处理统计判决问题. 往后我们将讨论参数 θ 被赋予先验分布的情况. 这称为 Bayes 判决问题.

（三）判决函数及其风险函数

设给定了一统计判决问题的三要素：取值于样本空间 $\mathscr{X}$ 内的样本 X 及其分布族 $\{F_\theta, \theta\in\Theta\}$，行动空间 $\mathscr{D}$ 以及损失函数

$L(d, \theta)$. 问题是要根据样本的值 x 去确定一个行动 d. 先不谈选择的优良性问题，很清楚，我们需要的是这样一个规则，根据它，就可以每当有了样本值 x 时，去确定一个行动 d. 换言之，我们需要的是一个定义于样本空间 $\mathscr{X}$ 内而取值于行动空间 $\mathscr{D}$ 内的函数. 这种函数称为**判决函数**或**决策函数**.

例如，在前面提到的商店从工厂进货的那个问题中，损失函数为(5.16). 一个可以考虑的判决函数是

$$\delta(X)=d_1, \text{当 } X/n\leqslant 1/6; \ \delta(X)=d_2, \text{当 } X/n>1/6. \tag{5.17}$$

形式上说，任何一个判决函数 $\delta(x)$ 都可以作为所给的统计判决问题的解. 为比较其优劣，就要看在使用各种判决函数时，其所造成的损失如何. 设样本为 x，则按判决函数 δ，应采取行动 $\delta(x)$. 若参数为 θ，则由此造成的损失为 $L(\delta(x), \theta)$. 这个量与样本值 x 有关，因而也是随机的. 考虑到样本分布为 $F_\theta(x)$，知采用判决函数 δ 而参数为 θ 时，平均损失为

$$\begin{aligned} R(\delta, \theta) &= E[L(\delta(X), \theta)] \\ &= \int_x L(\delta(x), \theta)dF_\theta(x). \end{aligned} \tag{5.18}$$

$R(\delta, \theta)$ 称为判决函数 δ 当参数取值 θ 时的**风险**. 而 $R(\delta, \theta)$ 作为 θ 在 Θ 上的函数，则称为 δ 的**风险函数**.

例 5.17 再考虑例 5.16，取判决函数(5.17)，来计算其风险函数. 当 $X\leqslant n/6$ 时，损失为 $10Np$. 当 $X>n/6$ 时，损失为 $2N(1-p)$. 于是按定义(5.18)，得

$$\begin{aligned} R(\delta, p) = {} & 10Np\sum_{i=0}^{[n/6]}\binom{n}{i}p^i(1-p)^{n-i} \\ & +2N(1-p)\sum_{i=[n/6]+1}^{n}\binom{n}{i}p^i(1-p)^{n-i}. \end{aligned}$$

按照 Wald 的理论，评价一个判决函数的唯一依据，就是其风险函数. 风险函数愈小愈好. 因此，若存在这样一个判决函数 δ^*，使对任何判决函数 δ 都有

$$R(\delta^*, \theta)\leqslant R(\delta, \theta), \quad \text{对任何 } \theta\in\Theta, \tag{5.19}$$

则称 δ^* 为此判决问题的**一致最优解**(一致是指(5.19)式对一切 $\theta\in\Theta$ 都成立). 若一致最优解存在, 则毫无疑问应采用它. 但是, 除了在某些没有现实意义的问题以外, 一致最优解不存在. 因此我们必须把标准放宽些, 即需要引进比一致最优准则更弱的优良准则. 我们将在以后几段中讨论这个问题.

(四) 统计推断与统计判决

统计推断和统计判决有什么差别和联系? 在有的统计著作中不大强调这个差别. 如在 Wald 的前引著作中, 把引进判决理论的目的视为将各种形式不同的统计推断问题用统一的方法去处理. 也有不少学者强调二者的差别, 认为不同之根本点在于: 判决问题要考虑行动的损失, 而统计推断问题, 尽管可以把种种具体推断说成是行动, 但不考虑行动的损失问题. 有的极端的意见认为推断与判决完全是两回事. 本书作者采取中庸的看法, 认为不能忽视二者的差别, 但也要承认, 二者确有密切联系.

统计推断是由样本 X 推断未知参数 θ. 此处"推断"一词是一种笼统的说法, 意指弄清楚与 θ 有关的某种情况. 例如, θ 等于多少? (估计问题) θ 是否 $\leqslant 0$? (检验问题)等等. 按流行的说法, 这是一种探求未知, 认识真理的科学研究性质的工作. 研究者力求使其结论尽可能与实际符合, 但他并不为可能的错误担当经济或其他后果. 例如, 通过用种种方法作试验估计光的速度 c. 我们力求使估计的误差小, 但当估计有较大误差时, 并不会"受罚". 总之, 在统计推断问题中, 我们努力想出一些办法使推断尽可能正确而不去考虑损失问题. 例如极大似然估计法. 我们从"似然性"的角度分析, 觉得这是一个可能给出良好估计的方法, 因此就采用它. 至于这个估计是某在某一问题中真能使某种意义下的损失尽可能小, 则不予考虑. 对似然比检验也可这样了解.

统计判决问题则不然. 虽则从某种意义上说, 也可以讲判决的任务是通过样本 X 弄清 θ 的有关情况——因为, 若已知 θ, 则由损失函数的形状可定出最优行动. 所以初看起来, 似乎判决问

题的解决可以分两步走：第一步用 X 对 θ 作出估计 $\hat{\theta}$. 第二步找一个行动 d，使 $L(d, \hat{\theta})$ 达到最小. 其实不然，这样做出的解通常不能使风险函数在一定意义下达到最小. 因此，我们打一开始就须盯着损失. 举一个形象的例子. 在点估计中，无偏性可以被认为具有其优点. 但在判决问题中，正偏差与负偏差的后果可以很不一样. 比方说，少生产了少得些利润，损失小些；多生产了造成积压，损失大些. 在这种情况下，对市场容量的一个略为偏低的估计，也许要比无偏估计为优. 另外，由于不同的当事者所具有的条件不同，同一个行动所造成的后果也可以不一样. 例如仓库设备较紧的厂，在发生积压时要租用他人的仓库而付出高额费用，损失就更大些. 因此，同一个解可能适用于甲而不适用于乙. 在推断问题中不应有这种情况，因为推断的任务是探求真理，这对于甲乙都是一视同仁的.

推断与判决的这个差别产生一个看法，认为在推断中不应参入主观成分，如主观先验分布之类的东西. 而判决问题的解，因与当事者的条件和所掌握的情报有关，故恰当的解不可能也不应排除主观因素. 这个观点就一般而论无疑是合理的. 问题在于：怎样的推断可以被认为是客观的，甚至是否存在这种推判，都是有争议的问题，例如，前面已提到过，有些表面上不涉及先验分布的解，实际上却有着一个潜在的且甚不合理的先验分布.

至于统计判决问题与推断问题的联系，表面上看，在于任何推断都可以赋予行动的意义. 实质上，联系在于：**统计判决的方法可以作为产生优良推断的一种手段**. 例如，“无偏估计的方差愈小愈好”这一准则，表面上固然可以从纯推断的观点去解释，但实质上，它无非是在平方损失（即损失函数为 $L(d, \theta)=(\theta-d)^2$）之下的最小风险的估计. 当采用种种不同的损失函数时，就可以得到种种推断方法. 在有些问题中，对一类广泛的损失函数来说，最优推断是一致的. 这样，该推断方法的优良性就得到更多方面的支持. 例如，设 $X_1, \cdots, X_n \sim N(\theta, 1)$，为估计 θ，若限制估计为无偏的，则不仅在平方损失下，且对广大的一类损失函数，样本均值 $\bar{X}$ 都是

最小风险估计．这样，$\overline{X}$ 作为 θ 的估计的优良性就得到更多的支持。

总之，尽管在推断问题中不涉及到当事者的损失，但这并不排斥把损失这个概念作为产生优良统计推断方法的一种手段．这样一来，由于损失函数的灵活性，就可以提出大量的优良性问题，从而丰富了统计推断的理论和方法。

例如，拿假设检验问题来说，从纯推断的观点，就是按照R. A. Fisher 的作法，对具体的样本 x，去考察它对原假设或对立假设提供了多大的支持，这不涉及任何损失的考虑．反之，也可以从统计判决的观点，规定当决定正确时损失为 0，而当决定错误时损失为 1，分别以 d_1, d_2 记接受和否定原假设 $\theta\in\Theta_H$（对立假设为 $\theta\in\Theta_K$）的决定，并以 $\beta_\varphi(\theta)$ 记检验 φ 的功效函数，则作为一个判决函数 φ，其风险函数为

$$R(\varphi,\ \theta)=\begin{cases}\beta_\varphi(\theta), & \text{当 } \theta\in\Theta_H;\\ 1-\beta_\varphi(\theta), & \text{当 } \theta\in\Theta_K.\end{cases}$$

于是，Neyman-Pearson 关于控制第一类错误概率的原则，可以用统计判决的语言描述为：指定 $\alpha\in(0,\ 1)$，把风险函数 $R(\varphi,\ \theta)$在 Θ_H 上的值控制在不超过 α，而使 $R(\varphi,\ \theta)$在 Θ_K 上尽可能小．我们之所以说 Neyman-Pearson 理论包含了 Wald 理论的一些概念，其道理可以由这里看出来．而且，由此我们更进一步领悟到，Fisher 和 Neyman-Pearson 关于假设检验理论差别之所在，是纯推断观点与判决理论观点的差异．

对 Neyman 的置信区间理论也可作类似的处理．对区间估计问题而言，可以把行动空间 $\mathscr{D}$ 定义为一切区间$[a,\ b]$的集．我们可定义两个损失函数 $L_1([a,\ b],\ \theta)$和 $L_2([a,\ b],\ \theta)$，其中

$$L_1([a,\ b],\ \theta)=\begin{cases}0, & \text{当 } a\leqslant\theta\leqslant b;\\ 1, & \text{其他情况}.\end{cases}$$

而损失 L_2 则反映区间的精度，例如，可定义为

$$L_2([a,\ b],\ \theta)=b-a.$$

这时，对任一区间估计 $\delta(x)=[A(x),\ B(x)]$，第一个损失 L_1 导

致风险函数

$$R_1(\delta,\ \theta)=1-P_\theta(A(X)\leqslant\theta\leqslant B(X)).$$

而第二个损失 L_2 的风险则是区间 $\delta(x)$ 的平均长度:

$$R_2(\delta,\ \theta)=E_\theta[B(X)-A(X)].$$

指定置信系数 $1-\alpha$ 可理解为要求 $R_1(\delta,\ \theta)\leqslant\alpha$, 然后在这个条件下要求 $R_2(\delta,\ \theta)$尽可能小. 这样,纯推断性质的 Neyman 区间估计理论, 也可以用判决问题的观点去描述. 由于损失函数的灵活性, 我们可以提出多种多样的问题. 例如, 可以只引进一个损失函数

$$L([a,\ b],\ \theta)=|a-\theta|+|b-\theta|.$$

这个损失同时兼顾了估计$[a,\ b]$的可靠性和精确性的方面.

不言而喻,存在着下面的问题: 取怎样的损失函数最可能导致良好的推断?如果问题确是纯推断性的,而对"良好推断"一词又无确切的定义,这问题是无法解决的.

(五) Bayes 准则

前已说过,一致最优的判决函数通常不存在,故有必要引进某种限制较宽的优良性准则. 如果我们考虑参数 θ 有某种先验分布 $H(\theta)$, 就可以自然地引出一个优良性准则. 那就是把风险函数 $R(\delta,\ \theta)$对 θ 再取一次平均, 得

$$R_H(\delta)=E^\theta[R(\delta,\ \theta)]=\int_\Theta R(\delta,\ \theta)dH(\theta), \qquad (5.20)$$

其中 E^θ 表示期望值是对 θ 求的. $R(\delta)$称为判决函数 δ 在先验分布 H 之下的 Bayes 风险. Bayes 风险达到最小的判决函数称为判决问题的 Bayes 解.

定义 5.3 设判决函数 δ 的风险函数记为 $R(\delta,\ \theta)$, 而在先验分布 H 之下的 Bayes 风险 $R_H(\delta)$由(5.20)式定义. 设 δ^* 为一判决函数. 若对任何判决函数 δ 都有 $R_H(\delta^*)\leqslant R_H(\delta)$, 则称 δ^* 为统计判决问题的一个 Bayes 解, 或者说, δ^* 是一个 Bayes 判决函数.

我们注意到，(5.20)是风险函数 $R(\delta, \theta)$ 以先验分布 H 为权的一种加权平均. 从这个角度看，我们可以完全撇开 Bayes 学派的观点，而直接把 H 解释为一种加权因子，把 $R_H(\delta)$ 视为衡量 δ 的优良性的一项综合指标. 这样，Bayes 准则可以完全从频率学派的观点得到解释. 在这里，Bayes 统计的先验分布概念只是起了这样的作用：当有理由选择 H 作为 θ 的先验分布时，H 是一个良好的权. Bayes 判决函数容许频率解释(或更确切地说，不须从先验分布的角度去解释)这一点，是 Bayes 判决问题和 Bayes 推断问题根本差别之所在.

往下考虑求 Bayes 解的问题. 以 $H(\theta|x)$ 记当样本值为 x 时，θ 的后验分布. 又设 $d\in\mathscr{D}$ 为一行动.

定义 5.4 设 $L(d, \theta)$ 为损失函数，则称

$$R(d|x)=\int_{\Theta} L(d, \theta) H(d\theta|x) \tag{5.21}$$

为当得到样本 x 时，**行动 d 的后验风险**. 换句话说，后验风险就是 $L(d, \theta)$ 作为 θ 的函数在后验分布之下的期望值.

下面的定理称为**后验风险最小原则**，它告诉我们怎样去寻求 Bayes 判决函数.

定理 5.1(后验风险最小原则) 对任何样本值 x，若存在行动 d_x 使后验风险达到最小，即

$$R(d_x|x)=\min_{d\in\mathscr{D}} R(d|x), \tag{5.22}$$

则由下式所定义的判决函数 δ_H：

$$\delta_H(x)=d_x, \quad 一切\ x\in\mathscr{X} \tag{5.23}$$

是一个 Bayes 判决函数.

证 设 δ 为任一判决函数，则由定义，知 δ 的 Bayes 风险为 $R_H(\delta)=E^{(X,\theta)}[L(\delta(X), \theta)]$. $E^{(X,\theta)}$ 表示求期望值时，把 X 和 θ 都看成随机的. 现分两步计算 $R_H(\delta)$：第一步给定 $X=x$，在这个条件下计算 $L(\delta(X), \theta)$ 的条件期望值. 由于在给定 $X=x$ 时 θ 的条件分布为 $H(\theta|x)$，知这个条件期望值为

$$\int_{\Theta} L(\delta(x),\ \theta) H(d\theta|x) = R(\delta(x)|x). \tag{5.24}$$

第二步再对 x 求期望. 若以 Q 记 X 的边缘分布(参看(二)), 则有

$$R_H(\delta) = \int_{\mathscr{X}} R(\delta(x)|x) dQ(x). \tag{5.25}$$

由(5.21)~(5.24), 知对任何 $x \in \mathscr{X}$, 有

$$R(\delta_H(x)|x) = \min_{d \in \mathscr{D}} R(d|x) \leqslant R(\delta(x)|x).$$

于是得到

$$\begin{aligned} R_H(\delta_H) &= \int_{\mathscr{X}} R(\delta_H(x)|x) dQ(x) \\ &\leqslant \int_{\mathscr{X}} R(\delta(x)|x) dQ(x) = R_H(\delta). \end{aligned}$$

这证明了本定理.

例 5.18 设问题为估计参数 θ, 并设损失为平方: $L(d,\ \theta) = (\theta - d)^2$. 分别以 $m(x)$ 和 $\sigma^2(x)$ 记后验分布 $H(\theta|x)$ 的均值和方差, 则有

$$R(d|x) = [d - m(x)]^2 + \sigma^2(x). \tag{5.26}$$

因此立见 Bayes 判决函数为

$$\delta_H(x) = m(x), \tag{5.27}$$

其 Bayes 风险为

$$R_H(\delta_H) = \int_{\mathscr{X}} \sigma^2(x) dQ(x). \tag{5.28}$$

Q 为 X 的边缘分布.

就是说, 以后验分布均值去估计 θ, 是本问题的 Bayes 解. 在§5.1第(四)段中讨论 Bayes 推断时, 也曾把这个估计提出来. 但在该处, (5.27)只是可供选择的推断之一, 而在此处则是唯一的 Bayes 解.

例 5.19 设参数 θ 有两个可能值 $\theta_1,\ \theta_2$. 样本分布有概率函数 $f(x,\ \theta)$. 考虑检验问题 H: $\theta = \theta_1 \leftrightarrow K$: $\theta = \theta_2$. 损失函数定义为(d_1: 接受 H. d_2: 否定 H)

$$L(d_1,\ \theta_1) = L(d_2,\ \theta_2) = 0,\ L(d_1,\ \theta_2) = a,\ L(d_2,\ \theta_1) = b.$$

又设 θ 的先验分布 H 为：$H(\theta_1)=p$, $H(\theta_2)=1-p$. 当得到样本 x 时，θ 的后验分布为

$$H(\theta_1|x)=pf(x,\ \theta_1)/[pf(x,\ \theta_1)+(1-p)f(x,\ \theta_2)],$$

$$H(\theta_2|x)=(1-p)f(x,\ \theta_2)/[pf(x,\ \theta_1)+(1-p)f(x,\ \theta_2)].$$

于是行动 d_1, d_2 的后验风险分别为

$$R(d_1|x)=a(1-p)f(x,\ \theta_2)/[pf(x,\ \theta_1)+(1-p)f(x,\ \theta_2)],$$

$$R(d_2|x)=bpf(x,\ \theta_1)/[pf(x,\ \theta_1)+(1-p)f(x,\ \theta_2)].$$

按定理 5.1，知此问题的 Bayes 解为

$$\delta_H(x)=\begin{cases}d_1, & \text{当 } f(x,\ \theta_2)/f(x,\ \theta_1)\leqslant a(1-p)/bp;\\ d_2, & \text{当 } f(x,\ \theta_2)/f(x,\ \theta_1)>a(1-p)/bp.\end{cases} \tag{5.29}$$

这个解的形式与 Neyman-Pearson 基本引理所提供的 UMP 检验符合．不同的是在 Neyman-Pearson 理论中，临界值由给定的检验水平 α 确定(且还要用到 $f(X,\theta_2)/f(X,\theta_1)$ 在 $\theta=\theta_1$ 之下的分布)，而在此处则取决于 a, b, p. a、b 值的差异反映了两类错误后果之不同，这自然会有使我们更倾向于接受或者否定 H. 因此，a、b 值的选择正好起着水平 α 的选择的作用． 在此又一次看到 Neyman-Pearson 理论与判决函数的联系．

在有的情况下，不论对那个判决函数 δ，总有 $R_H(\delta)=\infty$. 这时，任一个判决函数 δ 都是 Bayes 判决函数，在这个情况下，Bayes 准则丧失了意义．但是，后验风险最小的解仍可能是唯一的． 从 Bayes 风险的角度看，这个解固然与其他任何解一样，但是从其他角度看则可能有其优点，因此，常把在定理 5.1 中确定的后验风险最小的那个判决函数，称为**推广意义下的 Bayes 判决函数**.

另外，在 $R_H(\delta)$的定义(5.20)式中，$H(\theta)$不一定是概率分布，也可以是广义的概率分布，即可以有 $\int_\Theta dH(\theta)=\infty$. 若把 $H(\theta)$ 看作权而不赋予先验分布的意义，则这与 $\int_\Theta dH(\theta)=1$ 的情况比并无本质不同，而我们仍可提出 $R_H(\delta)$最小的准则．特别是，如果象 § 5.1(二)中那样，由 H 所确定的“后验分布”仍确是真正的概

率分布，则定理5.1仍有效．当然，也存在这样的问题：对任何δ，有$\int_{\Theta} R(\delta, \theta) dH(\theta)=\infty$．这时，与$H$为真正的概率分布的情况一样，由定理5.1确定的解仍可能有某种优越性，可称之为推广意义下的Bayes解．

例5.20 设$X_1, \cdots, X_n \sim N(\theta, 1)$，要估计$\theta$．损失函数为$L(d, \theta)=(\theta-d)^2$，先验分布为广义的：有密度$h(\theta)\equiv 1$．据§5.1（二），当有了样本值$x_1, \cdots, x_n$时，$\theta$的“后验分布”为正态$N(\bar{x}, 1/n)$，其均值为$\bar{x}$．因为损失函数为$L(d, \theta)=(d-\theta)^2$，知Bayes解为

$$\delta_H(X)=\bar{X}. \tag{5.30}$$

此解的风险函数为$E_\theta(\bar{X}-\theta)^2=1/n$，而Bayes风险为

$$R_H(\delta_H)=\int_{-\infty}^{\infty} \frac{1}{n} d\theta=\infty.$$

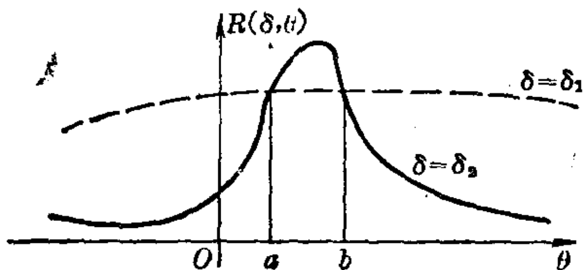
因此，对一切δ有$R_H(\delta)=\infty$．故在本例中，Bayes准则无意义．但是，由定理5.1只给出唯一解(5.30)，而这个解也确有其优点．

（六）Minimax准则

仍以$R(\delta, \theta)$记判决函数δ的风险函数，Θ为参数空间．则当采用判决函数δ时，使用者可能遭受的最大风险为

$$M(\delta)=\max_{\theta \in \Theta} R(\delta, \theta). \tag{5.31}$$

在有些问题中，当事者担心出现最坏的情况，其时风险大到超出所能承担的程度．这时他有兴趣于考虑这样的判决函数δ，其$M(\delta)$能尽可能小．这导致以下的准则，通常称为**Minimax准则或极小化极大准则**：



定义5.5 设δ^*为一判决函数．若对任何判决函数δ，都有$M(\delta) \geqslant M(\delta^*)$，则称$\delta^*$为此统计判决问题的**Minimax解**，也称$\delta^*$为**Minimax判决函数**．当问题为估计

或检验时，也称 δ^* 为 **Minimax** 估计或 **Minimax** 检验.

关于 Minimax 解的研究主要集中在点估计，在这方面得出了不少的深入结果. Minimax 准则一般是一个比较保守的准则. 形象地说，它"不求得到很多，但求不失去很多"，常有如上页图这种情况，其中 $M(\delta_1)<M(\delta_2)$. 故按Minimax 准则而言，δ_1 优于 δ_2. 但是通观二者的风险函数，发现对大多数 θ 值而言，δ_2 优于 δ_1. 如果我们没有足够的先验信息认为 θ 处在 a, b 之间，就很难说 δ_1 一定优于 δ_2 了. 因此，Bayes 学派认为，只是在人们对于 θ 的先验分布很没有把握的时候，作为一种替代，才使用 Minimax 解. 只要对先验分布有一定的把握，则宁肯用 Bayes 准则. 按 Bayes 学派的观点，Bayes 准则是一个较富进取性的准则，因为它力图最大限度地使用已有信息去争取最好的结果.

Minimax 准则的一个不方便之处，在于求 Minimax 解通常比较困难. 对若干重要的估计问题及平方损失函数，已找到了 Minimax 解. 它们大都是通过定理 5.2 和 5.3 中提供的方法获得的. 而且，这两个定理与其说是求 Minimax 解的方法，不如说是验证某一特定的解为 Minimax 解的方法.

定理 5.2 设 δ^* 为对某个先验分布 H 的 Bayes 解，且 δ^* 的风险函数恒等于一个有限常数 c: $R(\delta^*, \theta)=c$ 对任何 $\theta\in\Theta_H$，则 δ^* 为一个 Minimax 解.

证明很简单. 若 δ^* 不是 Minimax 解，则将存在判决函数 δ，使 $M(\delta)=c'<c$. 这时将有

$$R_H(\delta)=\int_\Theta R(\delta, \theta)dH(\theta)\leqslant\int_\Theta M(\delta)dH(\theta)$$

$$=c'<c=R_H(\delta^*).$$

这与 δ^* 为 Minimax 解相矛盾.

例 5.21 设 $X\sim B(n, p)$，要估计 p. 损失函数为 $(d-p)^2$，要求 p 的 Minimax 估计.

取 p 的共轭先验分布 $\mathrm{Be}(a, b)$，则当样本 X 取值 x 时，p 的后验分布为 $\mathrm{Be}(x+a, n+b-x)$. 由于损失函数为平方 $(d-p)^2$,

由例 5.18 知 Bayes 估计为此 Beta 分布的期望值，此值易算出为 $\delta_{ab}(x)=\frac{a+x}{n+a+b}$. 其风险函数为

$$R(\delta_{ab}, p)=E_p\left[\frac{X+a}{n+a+b}-p\right]^2$$
$$=\mathrm{Var}_p(X/(n+a+b))$$
$$+\left\{\frac{1}{n+a+b}(E_p(X)+a)-p\right\}^2.$$

考虑到 $E_p(X)=np$, $\mathrm{Var}_p(X)=np(1-p)$, 易算出

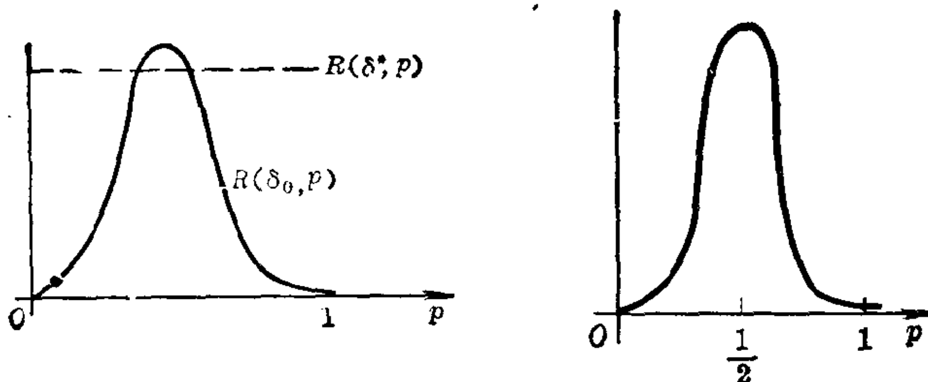
$$R(\delta_{ab}, p)=(n+a+b)^{-2}\{np(1-p)+[a-(a+b)p]^2\}.$$

若取 $a=b=\sqrt{n}/2$, 则易见上式右边恒等于一个常数 $n/[4(n+\sqrt{n})^2]$. 于是按定理 5.1, 知

$$\delta_{\sqrt{n}/2,\sqrt{n}/2}(x)=\frac{x+\sqrt{n}/2}{n+\sqrt{n}} \tag{5.32}$$

是 p 的 Minimax 估计.

由定理 5.1, 并通过这个实例, 我们注意到以下几点: 1° Minimax 准则从其形式到内容都与先验分布无关, 是频率学派也能接受的一个概念. 然而, 在一些情况下, 由这个准则提供的解却是在一定先验分布下的 Bayes 解. 2° 也可以这样看: 本例中引进的先验分布 Beta$(\sqrt{n}/2, \sqrt{n}/2)$, 完全不过是一个解题的工具, 不必赋予它以 Bayes 学派的那种意义. 因此, 人们可以对 Bayes 学派的基本观点持异议, 但不能否定其作用. 对于这类应用(即作为解题或论证的手段), 频率学派自然也是不拒绝的. 3° 传统的估计 $\delta_0=\overline{X}$ 的风险函数为 $R(\delta_0, p)=p(1-p)/n$. 此判决函数与 δ^* 的风险函数如下页左图所示. 由图看出, $M(\delta_0)>M(\delta^*)$, 故 δ_0 非 Minimax 解. 但是对多数 p 值, $R(\delta_0, p)$ 比 $R(\delta^*, p)$ 小, 尤以当 n 大时为然. 因此, 除非有某种特殊的原因, 人们一般仍宁肯用 δ_0 而不用 δ^*. 4°. Minimax 解 δ^* 所相应的先验分布为 Beta 分布 Beta$(\sqrt{n}/2, \sqrt{n}/2)$, 其密度函数的图形如下页右图所示. 此函数关于 $p=1/2$ 对称, 在 $p=1/2$ 的两边迅速衰减, 尤以 n 大时为然. 所以, 在 n 大时, 先验分布的概率极大地集中在点 1/2 的附近.



如果我们所有的关于 p 的先验知识并未提供这种证据(即先验概率集中在点1/2的附近),甚至是其反面(例如,废品率 p 一般较小,其先验分布的概率应集中在 $p=0$ 附近),则我们没有根据认为,Minimax 解(5.32)是合理的. 这就是我们前面在§5.1(六)段中提到的 Bayes 学派的那个论点:有些表面上不涉及先验分布的解,事实上隐含着甚不合理的先验分布.

定理 5.2 的使用面较窄,因一般很难找到一个其风险函数为常数的 Bayes 解. 下面的定理的应用要广得多.

定理 5.3 设一个统计判决问题在先验分布 H_k 之下的 Bayes 解为 δ_k, δ_k 的 Bayes 风险为 r_k, $k=1, 2, \cdots$. 假定

$$\lim_{k\to\infty} r_k = r < \infty, \tag{5.33}$$

又设 δ^* 为一判决函数,满足条件

$$M(\delta^*) \leqslant r, \tag{5.34}$$

则 δ^* 为此判决问题的 Minimax 解.

证 仍用反证法. 设 δ^* 非 Minimax 解,则存在判决函数 δ,使 $M(\delta) < M(\delta^*)$. 则由(5.33),(5.34),知当 k 充分大时,有 $M(\delta) < r_k$. 于是 δ 在先验分布 H_k 之下的 Bayes 风险 $R_{H_k}(\delta)$ 满足

$$R_{H_k}(\delta) \leqslant M(\delta) < r_k = R_{H_k}(\delta_k).$$

这显然与 δ_k 是先验分布 H_k 下的 Bayes 解矛盾,因而证明了本定理.

例 5.22 设 $X_1, \cdots X_n \sim N(\theta, 1)$,损失函数为 $L(d, \theta) =$

$(d-\theta)^2$，求 θ 的 Minimax 估计.

找一串先验分布 $\{H_k\}$, $H_k=N(0, k^2)$, $k=1, 2, \cdots$. 由于损失函数为 $(d-\theta)^2$，依例 5.18，知 Bayes 解为后验分布的均值. 因先验分布为 $N(0, k^2)$，按例 5.2 的结果，知 $\delta_k(x)=nk^2\bar{x}/(1+nk^2)$. 其风险函数为

$$
\begin{aligned}
R(\delta_k, \theta) &= E_\theta\left(\frac{nk^2\overline{X}}{1+nk^2}-\theta\right) \\
&= \mathrm{Var}_\theta\left(\frac{nk^2\overline{X}}{1+nk^2}\right)+\left\{\frac{nk^2}{1+nk^2}E_\theta(\overline{X})-\theta\right\}^2 \\
&= \frac{nk^4}{(1+nk^2)^2}+\frac{\theta^2}{(1+nk^2)^2}.
\end{aligned}
$$

而 δ_k 的 Bayes 风险为：在 $\theta\sim N(0, k^2)$ 的条件下求上式右边的期望值，结果为

$$
r_k=R_{H_k}(\delta_k)=\frac{nk^4}{(1+nk^2)^2}+\frac{1+k^2}{(1+nk^2)^2},
$$

显然有 $r=\lim\limits_{k\to\infty}1/n$. 取 $\delta^*(x)=\bar{x}$，则 $R(\delta^*, \theta)\equiv\dfrac{1}{n}$. 依定理 5.3，证明了 $\overline{X}$ 是 θ 的 Minimax 估计（在平方损失函数下）.

以往我们曾证明过，单纯作为一种统计推断而言，有一些准则都导致这估计量 $\overline{X}$. 例如，$\overline{X}$ 是 θ 的矩估计、极大似然估计，UMVUE 等. 这里又证明了在 Minimax 准则下，$\overline{X}$ 也是解. 仅从推断的角度看，这个结果也是有意义的. 因为它从一个新的角度验证了 $\overline{X}$ 这个估计的优良性.

定理 5.3 可以在一些更复杂的问题中使用，困难之点在于怎样去看出那一串具有定理中的性质的 $\{H_k\}$. 在不少问题中，$\{H_k\}$ 是从共轭先验分布类中去选. 进一步讨论已超出了基础课的范围，不在此给出了.

（七）同变原理

上两段讨论的 Bayes 准则和 Minimax 准则，都是用某种方法制定一个优良性的综合指标（$R_H(\delta)$, $M(\delta)$ 等），以之作为比较的

标准. 另一类制定优良性准则的方法是：先对判决函数 δ 提出某种看来是合理的要求，只考虑满足这种要求的判决函数的类 $\mathscr{A}$，然后在 $\mathscr{A}$ 内寻求一致最优者. 无偏性就是这样一种要求. 本段将简单地介绍一下另一个这样的要求，即**同变性**要求.

我们从一个简单例子入手. 设要估计某物件的重量 a. 将它放在一架天平上称量 n 次，得样本 $X_1, \cdots, X_n$. 我们用某个估计量$\delta(X_1, \cdots, X_n)$(例如 $\overline{X}$)去估计 a，假定 $X_1, \cdots, X_n \sim N(a, \sigma^2)$.

如果我们把坐标原点移至 $-c$，则物件的重量成为 $c+a$，数据转换为$X_i'=X_i+c$, $i=1, \cdots, n$. 用估计量 δ，将得到估计值 $\delta(X_1+c, \cdots, X_n+c)$. 但这是 $c+a$ 的估计，还原为 a 的估计，应是

$$\delta(X_1+c, \cdots, X_n+c)-c. \qquad (5.35)$$

但若不移动原点，则将用 $\delta(X_1, \cdots, X_n)$估计 a. 于是，若我们提出看来是很合理的要求：估计值不应与度量原点的取法有关，则我们必须要求，所用的估计量 δ 满足条件

$$\delta(X_1+c, \cdots, X_n+c)=\delta(X_1, \cdots, X_n)+c, \text{一切实数 } c. \qquad (5.36)$$

这就是对 δ 在变换$\{X_i'=X_i+c,\ i=1, \cdots, n\}$之下的同变性要求. 满足这个要求的估计量，称为**同变估计量**(在给定的变换之下). 若引进了一定的损失函数，则可以去寻求在同变估计类中，其风险一致最小者. 它称为**最优同变估计**(在一定变换和一定损失之下). 在本例中还有另一个自然的变换，即$\{X_i'=cX_i,\ i=1, \cdots, n,\ 0<c<\infty\}$. 这相当于把度量单位由 1 改为 c^{-1} (例如，由公斤改为克). 相应的同变性要求为

$$\delta(cX_1, \cdots, cX_n)=c\delta(X_1, \cdots, X_n), \text{一切 } c>0. \qquad (5.37)$$

我们注意，常见估计 $\overline{X}$ 适合这两个变换下的同变要求.

这个例子所提出的概念可推广到一般的统计判决问题. 因为问题较为专门，在本课程中只能满足于大略介绍一下而不深入其细节. 一个统计判决问题要运用同变性，必须满足以下两个要求：

1. 能定义样本空间 $\mathscr{X}$ 到自身之上的一些一一对应的变换，它们构成一个群. 而且，这群内的每一变换都把样本分布变换到样本分布族内另一分布，而不能越出样本分布族.

例如，变换类 $\{X_i'=X_i+c,\ i=1,\ \cdots,\ n,\ -\infty<c<\infty\}$，是由样本空间 $\mathscr{X}$（即 n 维欧氏空间 R^n）到 $\mathscr{X}$ 上的一个变换群：$c=0$ 相应于单位变换. 以 $-c$ 代 c 得到逆变换，由 $c=c_1$ 和 c_2 产生的两个变换之积，等于由 $c=c_1+c_2$ 产生的变换. 又样本分布经变换后仍在正态族内：若 $X_i'=X_i+c,\ i=1,\ \cdots,\ n$，则 $X_1',\ \cdots,\ X_n'\sim N(a+c,\ \sigma^2)$.

经过样本的这一变换，样本分布族的参数也起了变换. 在本例中为 $a'=a+c$.

2. 损失函数与行动空间要有这样的性质：每当样本空间的一个变换 $X\to X'$（如上例的 $X_i'=X_i+c,\ i=1,\ \cdots,\ n$）引起参数 θ 的一个变换 $\theta\to\theta'$（如上例的 $a\to a+c$）时，在行动空间上可找到一个一对一的变换 $d\to d'$，使 $L(d',\ \theta')=L(d,\ \theta)$. 如在上例中，若损失函数为 $L(d,\ a)=(d-a)^2$，则可在行动空间中引进变换 $d'=d+c$. 这时有 $L(d',\ a')=(d'-a')^2=(d-a)^2$.

如果这些条件都满足了，就可以引进如下的定义（记号意义如上述）：如当 $\delta(x)=d$ 时，必有 $\delta(x')=d'$，则称判决函数 δ（在给定的变换群下）为同变的. 由此可见，同变性的使用是受到不少局限的. 不仅对分布族有要求，对损失函数也有其要求. 如在上例中，若取定损失函数为 $L(d,\ a)=\dfrac{(d-a)^2}{1+a^2}$，则同变性不可用. 一般，同变性的使用多限于参数为位置和刻度参数的情况.

我们不去细论同变性在统计判决中的种种应用和理论. 只举一个例子，就是在上述问题中，若取平方损失，则 $\overline{X}$ 是 a 的最优同变估计. 为此需要下面的有用的引理.

引理 5.1 设 $X_1,\ \cdots,\ X_n\sim N(a,\ \sigma^2)$，而 $f(X_1,\ \cdots,\ X_n)$ 满足条件："$f(X_1+c,\ \cdots,\ X_n+c)=f(X_1,\ \cdots,\ X_n)$ 对任何实数 c"，则 $\overline{X}$ 与 $f(X_1,\ \cdots,\ X_n)$ 相互独立.

事实上，作正交变换 $Y=(Y_1, \cdots, Y_n)'=Q(X_1, \cdots, X_n)'$，$Q$ 的第一行为 $\left(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}}\right)$. 则如以 q_{ij} 记 Q 的 (i,j) 元，将有 $\sum_{j=1}^{n} q_{ij}=0$, $i=2, \cdots, n$. 按引理 1.1，知 $Y_1, \cdots, Y_n$ 独立、$\overline{X}=Y_1/\sqrt{n}$ 只与 Y_1 有关. 将 $f(X_1, \cdots, X_n)$ 表为 $Y_1, \cdots, Y_n$ 的函数，设为 $f(X_1, \cdots, X_n)=g(Y_1, \cdots, Y_n)$. 当 $(X_1, \cdots, X_n)$ 变换为 $(X_1+c, \cdots, X_n+c)$ 时，$Y_1, \cdots, Y_n$ 分别变换为

$$Y_1'=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i+c)=Y_1+\sqrt{n}\,c,$$

$$Y_i'=\sum_{j=1}^{n} q_{ij}(X_j+c)=Y_i,\ i=2, \cdots, n.$$

于是，由 $f(X_1+c, \cdots, X_n+c)=f(X_1, \cdots, X_n)$，得

$$g(Y_1+\sqrt{n}\,c, Y_2, \cdots, Y_n)=g(Y_1, Y_2, \cdots, Y_n).$$

此式对一切实数 c 成立，这表明 $g(Y_1, \cdots, Y_n)$ 其实与 Y_1 无关，即有 $h(Y_2, \cdots, Y_n)$ 的形式. 故得

$$\overline{X}=Y_1/\sqrt{n},\ f(X_1, \cdots, X_n)=h(Y_2, \cdots, Y_n).$$

考虑到 $Y_1, \cdots, Y_n$ 相互独立，即知 $\overline{X}$ 与 $f(X_1, \cdots, X_n)$ 相互独立. 引理证毕.

现回到我们的问题，取 a 的任一同变估计 δ，并记 $\delta_0(X_1, \cdots, X_n)=\delta(X_1, \cdots, X_n)-\overline{X}$. 则由 δ 的同变条件(5.36)，知 δ_0 满足 $\delta_0(X_1+c, \cdots, X_n+c)=\delta_0(X_1, \cdots, X_n)$，对一切实数 c. 于是依引理 5.1，知 $\overline{X}$ 与 $\delta_0(X_1, \cdots, X_n)$ 独立. 又

$$\begin{aligned}R(\delta, a)&=E_a[\delta(X_1, \cdots, X_n)-a]^2\\&=E[(\overline{X}-a)+\delta_0(X_1, \cdots, X_n)]^2,\end{aligned}$$

由于 $\overline{X}$ 与 δ_0 独立，有

$$\begin{aligned}R(\delta, a)&=E_a(\overline{X}-a)^2+E_a[\delta_0^2(X_1, \cdots, X_n)]\\&\quad+2E_a(\overline{X}-a)E_a[\delta_0(X_1, \cdots, X_n)]\\&=E_a(\overline{X}-a)^2+E_a[\delta_0^2(X_1, \cdots, X_n)]\geqslant E_a(\overline{X}-a)\\&=R(\overline{X}, a),\end{aligned}$$

对一切 a 都成立. 于是证明了 $\overline{X}$ 是 a 的一切同变估计中风险一

致最小者(在平方损失下),即 $\overline{X}$ 是 a 的最优同变估计.

(八) 容许性

最后我们来谈谈由一致最优性的考虑所引起的一个概念. 这个概念本身并不是一个优良性准则,但在很大程度上可以说,它是任何优良判决函数所应具备的一个条件.

定义 5.6 仍以 $R(\delta, \theta)$记判决函数 δ 的风险函数. 若存在另一个判决函数 δ_1,使

1° 对任何 $\theta\in\Theta$,有 $R(\delta_1, \theta)\leqslant R(\delta, \theta)$.

2° 至少存在一个 $\theta_0\in\Theta$,使 $R(\delta_1, \theta_0)<R(\delta, \theta)$.

则称 δ_1 一致地优于 δ,而 δ 称为是**不可容许的**. 反之,若不存在一致地优于 δ 的 δ_1,则称 δ 是**可容许的**.

如果 δ_1 一致优于 δ, 则当采用 δ_1时, 所承担的风险总不超过使用 δ 时所承担的,且确在某些场合下风险更小. 因此,若把风险大小作为评价判决函数优良性的唯一标准, 则我们自没有理由舍 δ_1 而就 δ.就是说, δ 是不允许使用的. 这就是容许性一词的由来.

一个判决函数的容许性,取决于该判决问题的各要素,尤其是样本分布族和损失函数,因为这二者决定了风险. 另外,参数空间的范围也对容许性有影响. 例如, 设 $X_1, \cdots, X_n\sim N(\theta, 1)$,损失为$(d-\theta)^2$. 用 $\overline{X}$ 估计 θ. 若 θ 的范围允许在$(-\infty, \infty)$,则下面将证明它是容许的. 反之,若 θ 局限于 $a\leqslant\theta\leqslant b$, a, b 都是有限常数,则很明显,若定义估计量

$$\delta_1(X_1, \cdots, X_n)=\begin{cases}\overline{X}, & \text{当 } a\leqslant\overline{X}\leqslant b;\\ a, & \text{当 } \overline{X}<a;\\ b, & \text{当 } \overline{X}>b,\end{cases}$$

则 δ_1 将一致优于 $\overline{X}$. 这一点在直观上很明显. 数学上也很容易证明 $R(\delta_1, \theta)<R(\overline{X}, \theta)$对任何 $\theta\in[a, b]$. 我们把这一点留给读者作为练习.

容许性的问题就在于,在一给定的统计判决问题中,确定那些判决函数可容许或不可容许,这一点极难作到. 有时,可以讨论某

一特定类型的判决函数是否可容许，例如线性估计类，最有兴趣也是文献中研究最多的情况，是确定某一特定的判决函数是否可容许．这种判决函数一般都是有某种优良性，例如，是问题的 Minimax 解，或者是由直观方法而产生的．这方面的研究工作难度很大，因为缺乏一般的有效方法．有些情况可用下面的定理解决：

定理 5.4 设 δ_H 为在某先验分布之下的 Bayes 解．设参数空间 Θ 是 m 维欧氏空间 R^m 或其一部分，先验分布 H 及 Bayes 解 δ_H 满足如下的条件：

1° 对任何判决函数 δ，δ 的风险函数 $R(\delta,\theta)$ 是 θ 的连续函数．

2° 设 θ_0 为 Θ 中的任一点，$S_{\theta_0}(P)$ 为以 θ_0 为中心，半径为 $\rho>0$ 的开球体．则对任何 $\rho>0$，有 $H(\Theta\cap S_{\theta_0}(\rho))>0$.

3° δ_H 的 Bayes 风险 $R_H(\delta_H)$ 有限．

则 δ_H 是可容许的．

证 若不然，则存在判决函数 δ，使 $R(\delta,\theta)\leqslant R(\delta_H,\theta)$ 对一切 $\theta\in\Theta$，且存在 $\theta_0\in\Theta$ 使 $R(\delta_H,\theta_0)-R(\delta,\theta_0)=2\varepsilon>0$．因为 $R(\delta,\theta)$，$R(\delta_H,\theta)$ 都是 θ 的连续函数，故存在 $\rho>0$，使当 $\theta\in\Theta$ 而 $\|\theta-\theta_0\|\leqslant\rho$（$\|\theta-\theta_0\|$ 是 θ 和 θ_0 两点的欧氏距离）时有 $R(\delta_H,\theta)-R(\delta,\theta)\geqslant\varepsilon$．这时有

$$
\begin{aligned}
R_H(\delta)&=\int_\Theta R(\delta,\theta)dH(\theta)\\
&=\int_{\Theta_1}R(\delta,\theta)dH(\theta)+\int_{\Theta_2}R(\delta,\theta)dH(\theta).
\end{aligned}
$$

此处 $\Theta_1=\Theta\cap S_{\theta_0}(\rho)$，而 $\Theta_2=\Theta-\Theta_1$．因为在 Θ_1 上有 $R(\delta_H,\theta)\geqslant R(\delta,\theta)+\varepsilon$ 而在 Θ_2 上有 $R(\delta_H,\theta)\geqslant R(\delta,\theta)$，有

$$
\begin{aligned}
R_H(\delta_H)=\int_\Theta R(\delta_H,\theta)\,dH(\theta)&=\int_{\Theta_1}R(\delta_H,\theta)dH(\theta)\\
&+\int_{\Theta_2}R(\delta_H,\theta)\,dH(\theta)\geqslant\int_\Theta R(\delta,\theta)dH(\theta)\\
&+\varepsilon H(\Theta\cap S_{\theta_0}(\rho))+\int_{\Theta_2}R(\delta,\theta)dH(\theta)
\end{aligned}
$$

$$=R_H(\delta)+\varepsilon H(\Theta\cap S_{\theta_0}(\rho)).$$

由于 $\varepsilon>0$, $H(\Theta\cap S_{\theta_0}(\rho))>0$ 且 $R_H(\delta_H)<\infty$, 由上式知 $R_H(\delta_H)>R_H(\delta)$. 这与 δ_H 为先验分布 H 之下的 Bayes 解矛盾，因而证明了本定理.

例如，考虑例 5.21 中的 Miniman 估计(5.32). 对此例而言，先验分布 H 在 $0<p<1$ 内有处处大于 0 的密度，故条件 2° 满足，条件 3° 显然满足，至于条件 1°，只须注意任一估计量 δ 有风险函数

$$R(\delta, p)=\sum_{i=0}^{n}[\delta(i)-p]^2\binom{n}{i}p^i(1-p)^{n-i}.$$

它显然是 p 的连续函数. 因此定理 5.4 的条件都满足，而 Minimax 估计(5.32)为可容许的. 用这个方法不能证明常见估计 X/n 的容许性. 此估计确是可容许的，且可用下文所用的“C-R 不等式法”去证明之.

又如，设 $X_1, \cdots, X_n\sim N(\theta, 1)$, $-\infty<\theta<\infty$, 损失函数为 $(d-\theta)^2$. 设 c 为常数，$0<c<1$, 则 $c\overline{X}$ 是 θ 的可容许估计. 事实上，取先验分布 $\theta\sim N(0, \tau^2)$, 其中 τ 选择之使 $\frac{n\tau^2}{1+n\tau^2}=c$, 则由例 5.2 和例 5.18, 知 $c\overline{X}$ 是此先验分布之下的 Bayes 解. 不难验证，定理 5.4 的条件 1°～3° 都满足. 条件 1°、2° 显然，对条件 3°, 则须验证表达式

$$R(\delta, \theta)=\int_{-\infty}^{\infty}\cdots\int[\delta(x_1, \cdots, x_n)-\theta]^2\left(\frac{1}{\sqrt{2\pi}}\right)^n \times\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right]dx_1\cdots dx_n. \qquad (5.38)$$

当 $R(\delta, \theta)$ 处处有限(否则 δ 不能一致优于 $c\overline{X}$)时，可在积分号下对 θ 求导. 这个容易的细节留给读者.

令 $c\to 1$, 则得 $c\overline{X}\to\overline{X}$. 于是 $\overline{X}$ 是容许估计的极限. 可是，容许估计的极限不见得可容许，故这样还不能证明 $\overline{X}$ 是 θ 的可容许估计. $\overline{X}$ 的可容许性也不能由定理 5.4 的方法直接证明. 下面的方法基于 C-R 不等式，由它可以处理有关单参数指数族的参

数在平方损失之下的某些容许估计问题.

现设 δ 为 θ 的任一估计,满足条件

$$R(\delta, \theta) \leqslant R(\overline{X}, \theta) = \frac{1}{n}, \text{对任何实数 } \theta. \tag{5.39}$$

记 $E_\theta(\delta(X_1, \cdots, X_n)) = \theta + b(\theta)$, 则如上面指出的, $b'(\theta)$存在. 依 C-R 不等式,有

$$[1+b'(\theta)]^2/n + b^2(\theta) \leqslant 1/n, \text{一切 } \theta \in (-\infty, \infty). \tag{5.40}$$

由此式可得 $b'(\theta) \leqslant -lb^2(\theta)$, $l = n/2 > 0$. 有 $b'(\theta)/b^2(\theta) \leqslant -l$. 由(5.40)式知 $b'(\theta) \leqslant 0$ 对一切 θ, 故 $b(\theta)$为处处非增的. 现往证 $b(\theta) \geqslant 0$, 对一切 θ. 不然的话, 存在 θ_0 使 $b(\theta_0) < 0$. 由非增性, 知 $b(\theta) < 0$, 当 $\theta > \theta_0$. 故由 $b'(\theta)/b^2(\theta) \leqslant -l$ 知, 当 $\theta > \theta_0$ 时有

$$-l(\theta - \theta_0) \geqslant \int_{\theta_0}^{\theta} \frac{b'(\theta)}{b^2(\theta)} d\theta = \frac{1}{b(\theta_0)} - \frac{1}{b(\theta)}. \tag{5.41}$$

令 $\theta \to \infty$, 由此式将得 $b(\theta) \to 0$. 这与 $b(\theta_0) < 0$ 且 $b(\theta)$非增矛盾. 由此知 $b(\theta) \geqslant 0$, 对一切 θ. 现有两个情况:

1° $b(\theta) \equiv 0$. 这时 δ 为 θ 的无偏估计, 而 $\overline{X}$ 为 θ 的 UMVUE, 故有

$$R(\delta, \theta) = \mathrm{Var}_\theta(\delta) \geqslant \mathrm{Var}_\theta(\overline{X}) = R(\overline{X}, \theta).$$

故这时 δ 不一致优于 $\overline{X}$.

2° 存在 θ_0 使 $b(\theta_0) > 0$. 这时 $b(\theta) \geqslant b(\theta_0) > 0$, 当 $\theta < \theta_0$. 但当 $\theta \to -\infty$ 时, $b(\theta)$不能趋于 ∞, 否则(5.40)式不能对一切 θ 成立. 故 $\lim\limits_{\theta \to -\infty} b(\theta) = c$, $0 < c < \infty$. 因此, 在(5.41)式中改 θ_0 为 $\theta - 1$, 再令 $\theta \to -\infty$, 将得 $-l \geqslant \dfrac{1}{b(\theta-1)} - \dfrac{1}{b(\theta)} \to \dfrac{1}{c} - \dfrac{1}{c} = 0$, 这是不可能的. 因此当(5.40)成立时只能是情况 1°, 而一致优于 $\overline{X}$ 的估计 δ 不存在. 这证明了 $\overline{X}$ 的可容许性.

现设有 p 个总体, 其分布分别为 $N(\theta_1, 1), \cdots, N(\theta_p, 1)$. 从第 i 个总体中抽出样本 $X_{i1}, \cdots, X_{in} \sim N(\theta_i, 1)$, $i = 1, \cdots, p$, 且设合样本 $X_{11}, \cdots, X_{1n}, \cdots; X_{p1}, \cdots, X_{pn}$ 全体独立. 要估计 $\theta_1, \cdots, \theta_p$. 损失函数为: 当分别用 $d_1, \cdots, d_p$ 估计 $\theta_1, \cdots, \theta_p$ 时, 损失

为 $L((d_1, \cdots, d_p), (\theta_1, \cdots, \theta_p))=\sum_{i=1}^{p}(d_i-\theta_i)^2$. 一个自然的估计量是用 $(\bar{X}_1, \cdots, \bar{X}_p)$ 估计 $(\theta_1, \cdots, \theta_p)$, 其中 $\bar{X}_i=\sum_{j=1}^{n} X_{ij}/n$. 这估计在上述损失函数下是否可容许? 由于 $p=1$ 时我们已见到回答是肯定的. 因此, 自然地会猜测这一点对 $p>1$ 也成立. 可是不然, 对 $p=2$ 可证明上述估计确为容许的. 但当 $p\geqslant 3$ 时, C. Stein 在 1955 年一项工作中, 出人意外地证明了上述估计非容许. 这个结果引起了广泛的兴趣, 成为一系列研究的出发点. 它也启示我们: 容许性这个概念是一个很复杂的东西.

习　题

1. 设 N、n 都是自然数, $n\leqslant N$. (a) 证明

$$\sum_{m=0}^{N}\binom{m}{k}\binom{N-m}{n-k}=\binom{N+1}{n+1}, \quad k=0, 1, \cdots, n.$$

(提示: 把 $N+1$ 个球自左至右排成一列, 编上号 $1, \cdots, N+1$. 从中选取 $n+1$ 个, 选法有 $\binom{N+1}{n+1}$ 种、以 $1\leqslant i_1<i_2<\cdots<i_{n+1}\leqslant N+1$, 计算满足条件 "$i_{k+1}=m+1$" 的选法.) (b) 利用 (a) 解下述问题: 设样本 X 服从超几何分布: $P_M(X=x)=\binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n}$, $x=0, 1, \cdots$. 此处 N, n 已知而 M 为参数. 设 M 的先验分布为

$$P(M=k)=\frac{1}{N+1}, \quad k=0, 1, \cdots, N.$$

用后验分布均值估计 M. 证明此估计为 $\frac{(x+1)(N+2)}{n+2}-1$.

2. 设 $X_1, \cdots, X_n\sim R(0, \theta)$, $\theta>0$ 为参数. 设 θ 的先验分布为 $R(0, a)$, $a>0$ 已知. 用后验分布的均值估计 θ.

3. 设 $X_1, \cdots, X_n$ 为自一总体中抽出的独立随机样本, 总体密度为

$$f(x, \theta)=\begin{cases} e^{\theta-x}, & x>\theta; \\ 0, & x\leqslant\theta, \end{cases}$$

$-\infty<\theta<\infty$, θ 为参数, 其先验分布为 Cauchy 分布 $\left(\text{有密度 } \frac{1}{\pi(1+\theta^2)}\right)$. 求 θ 的广义极大似然估计.

4. 设 $X_1\sim N(\theta, 1)$ (样本大小为 1), θ 有广义先验密度 $h(\theta)=1$ 当 $\theta>$

0, $h(\theta)=0$ 当 $\theta\leqslant 0$. 证明: 用后验分布均值估计 θ 的结果为 $X_1+e^{-X_1^2/2}\Big/\int_{-X_1}^{\infty} e^{-y^2/2}dy$.

5. (a)设 $X_1, \cdots, X_m\sim N(a, 1)$, $Y_1, \cdots, Y_n\sim N(b, 1)$, a, b 为参数. 又合样本独立, 而 (a, b) 的先验分布为: a, b 独立, $a\sim N(\mu_1, \tau_1^2)$, $b\sim N(\mu_2, \tau_2^2)$. 检验假设 H: $a\leqslant b \leftrightarrow a>b$. (b)若第一、二类错误的损失分别为 c_1 和 c_2, 解上述问题.

6. 考虑第3题当 $n=1$ 的情况, 并将 θ 的先验密度改为: $h(\theta)=e^{-\theta}$ 当 $\theta>0$, $=0$ 当 $\theta\leqslant 0$. 检验假设 H: $\theta\leqslant 1\leftrightarrow K$: $\theta>1$.

7. 考虑第2题, 作 θ 的后验置信度为 $1-\alpha$ 的最短区间估计.

8. 设 X_1 为抽自指数分布族(1.8)的大小为1的样本, 而 θ 有广义先验密度如第4题, 作 θ 的后验置信度为 $1-\alpha$ 的最短区间估计.

9. 设 $X_1, \cdots, X_n\sim R(\theta_1, \theta_2)$, $\theta_1<\theta_2$ 为参数. (θ_1, θ_2)的先验分布定义为: θ_1 有指数密度 $e^{-\theta_1}$(当 $\theta_1>0$. 当 $\theta_1\leqslant 0$ 时密度为 0), 而在已知 θ_1 时, $\xi=\theta_2-\theta_1$ 的条件分布也有指数密度($e^{-\xi}$ 或 0, 视 $\xi>0$ 或 $\xi\leqslant 0$). 试求 θ_2/θ_1 的后验置信度为 $1-\alpha$ 的最短区间估计.

10. 试求第1题的估计的 Bayes 风险$\left(\text{结果为 } \dfrac{N(N+2)(N-n)}{6(n+2)}\right)$, 又: 计算常用估计 Nx/n 的 Bayes 风险, 并与上述 Bayes 风险比较谁大谁小.

11. 设 $X_1, \cdots, X_n\sim N(a, \sigma^2)$, a, σ^2 为参数. 先验分布为: $\dfrac{1}{2\sigma^2}$ 有指数分布(1.8)且 $\lambda=1$, 而在给定 σ 时, a 的条件分布为 $N(0, k\sigma^2)$, $k>0$ 已知, 要作 a 的区间估计. 且规定当用$[c, d]$去估计 a 时, 损失为

$$L(a, \sigma; c, d)=\begin{cases}\dfrac{d-c}{\sigma}, & \text{当} a\in[c, d];\\ \dfrac{d-c}{\sigma}+m, & \text{当} a\,\overline{\in}\,[c, d].\end{cases}$$

求此问题的 Bayes 解, 并考虑当 $k\to\infty$ 时的情况.

12. 设 $X\sim B(n, p)$. p 的先验分布为 $P(p=p_0)=P(p=1-p_0)=1/2$, $0<p_0<1/2$, p_0 已知. 又损失函数定为$(p-d)^2$. 求 p 的 Bayes 估计, 并证明: 可适当选择 p_0, 使此 Bayes 估计为 p 的 Minimax 估计(本题说明了(结合例): Minimax 估计可以由截然不同的先验分布导出, 且不同的先验分布可产生同一 Bayes 解).

13. 设 $X_1, \cdots, X_n\sim R(0, \theta)$, $\theta>0$. 要作 θ 的点估计, 损失函数定为 $L(\theta, d)=\left(\dfrac{d}{\theta}-1\right)^2$. 证明: 此问题在变换 $X_i'=cX_i$, $i=1, \cdots, n$, $c>0$ 之下

为同变，且最优同变估计有 $g_n\cdot\max\limits_{1\leqslant i\leqslant n} X_i$ 的形状，g_n 为一与 n 有关的常数．找出 g_n（提示：利用 $\max\limits_{1\leqslant i\leqslant n} X_i$ 为充分统计量）．

14．举例说明：矩估计和极大似然估计都可以是不可容许的．

15．设 $X_1, \cdots, X_n \sim N(\theta, 1)$，要估计 θ．损失函数为 $(\theta-d)^2$，取估计量 $\hat{\theta}_n = c_1X_1+\cdots+c_nX_n$．证明：若 $c_1+\cdots+c_n=1$，则除非 $c_1=\cdots=c_n=1/n$，$\hat{\theta}_n$ 是不可容许的．

16．同上题．以 m_n 记 $X_1, \cdots, X_n$ 的样本中位数．证明：m_n 不是 θ 的可容许估计，除非 $n\leqslant 2$（提示：m_n 为 θ 的无偏估计，而 $\bar{X}$ 为唯一的 UMVUE）．

17．设 X 为自 $N(\theta, 1)$ 中抽出的样本，θ 的先验分布为 $N(0, 1)$，要估计 θ．损失函数定为 $L(\theta, d)=e^{3\theta^2/4}(\theta-d)^2$．(a)证明：在后验风险最小的意义下，$\hat{\theta}=2X$ 是唯一的 Bayes 解．(b)证明 $\hat{\theta}$ 不可容许（提示：与 $\theta^*=X$ 比较）．

18．设 $X_1, \cdots, X_n$ 为自一总体中抽出的独立随机样本，总体密度为 $f(x, \theta)=e^{-x/\theta}/\theta$ 当 $x>0$，$f(x, \theta)=0$ 当 $x<0$．参数 $\theta>0$．取定损失函数 $(\theta-d)^2$．试用 C-R 不等式的方法，证明 $\bar{X}$ 是 θ 的可容许估计．

第六章　线性统计模型

本章集中讨论线性统计模型，简称为线性模型．线性模型在数理统计的理论和应用中都占据十分重要的地位．尤其在应用统计中，线性模型是人们所习惯使用的主要模型．

线性模型之所以有广泛应用，一方面是因为它概括了一大类实际统计问题，另一方面是因为它结构简单，处理比较方便，从而成为近似地处理其它类问题的较合适的模型．

如果从上世纪初 Gauss(1809)提出最小二乘法算起，线性模型已有相当悠久的历史，其理论也有了广泛深入的发展，并积累了许多行之有效的方法．因此，即使限于这个方向的一些主要论题，如回归分析，方差分析，都已有不少篇幅浩翰的专著予以阐述．本章所涉及的只能是这个方向的一些基本概念，基础理论和最常用的若干方法，这些仅是线性模型的入门知识．

本章和下一章将较多地使用矩阵工具．在某些理论探讨中，矩阵工具是必不可缺的，在许多推导和计算中，它将带来很大方便．这一工具已越来越为实际工作者所熟悉．本书所用到的矩阵和线性空间的基础理论，是理工科各系的高等数学课程中已经授过的．由于统计问题的特殊性，要用的有些结果可能在基础课程中未予强调．为了引用和参阅的方便，除了众所周知的浅显事实外，我们把要用的代数结果，基本上不加证明叙述在章末的附录A中．

§6.1　线性模型的概念和分类

(一)线性模型的概念

一个实际的统计问题，在进行数学的抽象之后，往往可以归结

为一个统计模型[1]. 这个模型一般地说，无非是联系着变量(包括随机变量和一般变量)和参数(已知的和未知的)的一组数学关系式. 作为数学对象的模型，当然不可能、也不必要完全等同于具体现象，但就我们所关心的那些方面而言，它应当能在一定的意义下有效地反映实际的统计问题. 尽管在建立统计模型的过程中，不免要涉及统计理论，但建立模型并不是统计学本身的任务. 所以，我们仅仅在必要时用极少篇幅提到模型的由来，而是把注意力集中在模型的分析和处理上.

假若我们要考察一个质点从时刻零到时刻 t 所走过的距离 s. 设质点作匀速直线运动，速度为 β，则有熟知的公式 $s=\beta t$ 作为这一现象的模型. 一般认为，时间 t 在一次观察或试验中是可以精确测定或严格控制的，不具有随机性，我们称这类变量为**可观察的一般变量**. 但在时刻 t 要测定 s 的值，就难免出现测量误差，于是描述上面现象的比较合理的模型应改为

$$s=\beta t+\varepsilon, \tag{6.1}$$

这里的 ε 表示测量误差. 测量误差的取值是随机性的，虽然它在每次测量中有一个客观值，却无法在实际中确定出来. 这一类量被称为**不可观察的随机变量**. s 此时也是随机变量，但在试验中可测得它的取值，称之为**可观察的随机变量**. 我们称可观察的随机变量在试验或观测中的取值为**观察值**. β 是一个未知参数，它在模型中有一个确定的未知值.

又若要考察两个随机变量 Y 和 X 的统计关系，这种关系一般不可能用函数关系来刻划，但在 X 给定的条件下，Y 与 X 的统计关系有可能具有(6.1)的形式. 这在后面要讲的线性回归中可以看出.

在这类例子和设想的基础上，抽象出线性模型的概念:

定义 6.1 设 Y 是可观察的随机变量，$x_1, \cdots, x_m$ 是可观察的

---

1) 第一章已经定义了统计模型. 这里将着重探讨其中具有特殊结构的某一类，即均值有线性结构或属于一个子空间的那一类.

一般变量，$\beta_1, \cdots, \beta_p$ 是未知参数，ε 是不可观察的随机变量，如果它们满足

$$Y=\sum_{j=1}^{p} f_j(x_1, \cdots, x_m)\beta_j+\varepsilon,\ f_j\text{是已知函数}, \tag{6.2}$$

则称(6.2)是**线性统计模型**[1]，简称为**线性模型**，(6.2)中的 ε 被称作**随机误差**，一般总假定 $E\varepsilon=0$，这是对线性模型的最少假定.

按照定义，(6.1)当然是线性模型，因测量误差被认为服从 $N(0,\sigma^2)$分布．并且，如设质点改作自由落体运动，那么，模型就变为

$$s=\gamma t^2+\varepsilon. \tag{6.3}$$

这里 γ 是未知参数．由于(6.3)对参数 γ 是线性的，按定义 6.1，它也是线性模型．从而要注意，线性模型的线性性，是对参数 $\beta_1, \cdots, \beta_p$ 而言的.

在线性模型(6.2)中，不妨将 $f_j(x_1, \cdots, x_m)$改记为 $\tilde{x}_j$，$j=1, \cdots, p$．从而可记

$$\sum_{j=1}^{p} f_j(x_1, \cdots, x_m)\beta_j \triangleq \sum_{j=1}^{p} \tilde{x}_j\beta_j.$$

于是，不失一般性可记线性模型为

$$Y=\sum_{j=1}^{p} x_j\beta_j+\varepsilon,\ E\varepsilon=0. \tag{6.4}$$

或者用均值的形式来表达，记(6.4)为

$$EY=\sum_{j=1}^{p} x_j\beta_j. \tag{6.5}$$

这就说明线性模型无非是一个随机变量其均值具有未知参数的线性结构的统计模型．在(6.4)的形式下，模型关于 $x_1, \cdots, x_p$ 也是线性的，但就定义而言，这一性质不是本质的.

在(6.5)中 EY 是 $x_1, \cdots, x_p$ 的函数，故有理由称 $x_1, \cdots, x_p$ 是**自变量**，并有理由借用函数关系中的名词，称 Y 为**因变量**．当然，我们不应把这里 Y 和 $x_1, \cdots, x_p$ 的统计依赖关系和函数关系相混淆，这里不存在决定性的因果关系.

---

1) 这里所下的定义是狭义的，但就本书而言是基本的．稍后可看到其推广情形.

(6.1)说明线性模型中测量误差可以是随机误差的一个来源．然而，在统计问题中，误差的另一个来源很值得重视，就其意义而言，或许更为重要．让我们来解释这类误差是如何引入模型的．例如要预报气温．影响气温的因素很多，倘若把影响微弱的因素也计算在内，不妨设有m个因素，记为$x_1, \cdots, x_m$．假设这些因素对于气温Y的影响可以用一个函数来表示，设为

$$Y=F(x_1, \cdots, x_m). \tag{6.6}$$

但从实践上看，把所有因素全考虑在内，或者根本做不到(如未认识到某项因素对Y是否有影响)，或者因花费太大不值得这样去做，还可能因为注意了细枝末叶反而使预报的效果变坏，所以，常常只考虑其中p个主要因素的影响(不妨记为$x_1, \cdots, x_p$)，而把次要因素的影响作为不可观察的随机变量对待．这样一种观点，从数学角度去看，似乎很不严格，但却是处理实际问题常用的办法，至于所得模型是否合用，尚需到实践中去检验和改进．在上述观点下(6.6)有可能改记为

$$Y=f(x_1, \cdots, x_p)+g(x_{p+1}, \cdots, x_m), \tag{6.7}$$

记$\varepsilon=g(x_{p+1}, \cdots, x_m)$．不失一般性，仍可设$E\varepsilon=0$．这是因为，如果$Eg(x_{p+1}, \cdots, x_m)=\beta_0\neq 0$，则可记$\varepsilon=g(x_{p+1}, \cdots, x_m)-\beta_0$，而在前一项$f(x_1, \cdots, x_p)$中加上$\beta_0$，从而仍使$E\varepsilon=0$．于是模型

$$Y=f(x_1, \cdots, x_p)+\varepsilon$$

中的误差ε就被认为来源于**方程近似**，即它是由方程不能精确反映实际问题而产生的．

在实际统计问题中，两类来源的误差同时出现的情形较为常见．从统计理论而言，并不追究误差的来源，而仅仅对它的分布特征感兴趣，如它的分布类型，均值，方差，以及各次观察中误差是否相互独立等．在后面的进一步讨论中，有时要对误差作进一步假定．诚然，对于试验工作者，误差来源可能是很重要的，尤其在出现系统误差($E\varepsilon\neq 0$)时，探究误差的来源就不容忽视．

在对模型(6.4)进行统计推断时，由于它含有p个未知参数和误差，常常需要将满足这一模型的试验，在不同的自变量值下进

行 n 次，以取得足够的试验数据．这里的试验次数 n 一般应大于 p．设第 α 次试验中自变量的取值为 $x_\alpha=(x_{\alpha1}, \cdots, x_{\alpha p})'$．称 x_α 是一个试验点．注意，我们用 α 这个角标专门表示试验的序次，这样可避免与变量的序次指标相混淆．相应的因变量 Y 的观察值记为 y_α, $\alpha=1, \cdots, n$．从而有样本满足的模型为

$$y_\alpha=\sum_{j=1}^{p} x_{\alpha j}\beta_j+\varepsilon_\alpha,\ E\varepsilon_\alpha=0,\ \alpha=1, \cdots, n. \tag{6.8}$$

记 $\boldsymbol{y}=(y_1, \cdots, y_n)'$, $\boldsymbol{X}=(x_{\alpha j})$, $\boldsymbol{\beta}=(\beta_1, \cdots, \beta_p)'$,

$\boldsymbol{\varepsilon}=(\varepsilon_1, \cdots, \varepsilon_n)'$,

(6.8)可简记为

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon},\ E\boldsymbol{\varepsilon}=\boldsymbol{0}. \tag{6.9}$$

这里的 $\boldsymbol{X}$ 通常称作设计矩阵，y 为观察值向量．

在一些教科书中，线性模型概念就是直接从(6.9)引入的．为区别起见，我们可称(6.9)为**样本模型**．今后在谈到线性模型时，我们将更多地从(6.9)出发，概念上常不加区分，相信不会引起混淆．

在具体统计问题中，将根据精确度要求(在一定的概率意义下表述)和取样所必须花费的代价来适当地选择试验次数 n．本章将限于讨论 n 固定时的小样本理论．线性模型的大样本理论可参看陈希孺《数理统计引论》．

(二)线性模型的分类

对于线性模型的分类，可以从各种角度进行，谈不上有很严格的标准．一种分类是从试验工作者的角度出发的，以模型的来源和要解决的实际问题作为分类的标准；另一种从统计工作者出发，则以模型的数学特征和它所侧重的统计问题为依据．我们倾向于后一种，并遵循已经形成的惯例．

让我们从自变量的性质谈起．如果模型中涉及的自变量的取值是长度、时间、体积、重量这一类可以连续变化的量，我们称这类自变量所代表的因子(或称因素)为**数量因子**．如考虑人的体重对

于身高的统计依赖，这里身高是一个数量因子．在另一些统计问题中，如农业试验，要考察影响小麦产量的因子，如土地、品种和肥料．土地和品种是按某一块、某一种来区分的，不是连续变化的量，称这类因子为**属性因子**．这类因子在统计中使用时需要数量化，但它的量值常常取1或0，表示出现与否，也称作标号(label)**变量**．实际上是把每块土地，每个品种各作为一个因子看待．就肥料而言，如果只考虑种类的区别，当然是属性因子．但若感兴趣于施肥量，似乎是一个数量因子了．可是，在试验中，不可能让施肥量连续变化去考察它的效应．我们只能就施肥量的若干个等级来考察，如区分10斤、20斤、30斤三个等级(称为**水平**)，实际上可以用三个标号变量来刻划，这种做法叫作将数量因子属性化．

根据自变量因子的性质，可将线性模型分为三类：凡自变量因子都是数量因子，就称这个模型是**回归分析模型**；如果自变量因子均为属性因子，则称此模型为**方差分析模型**；倘若自变量因子中，既有属性因子，也有数量因子，就称之为**协方差分析模型**．

另一种公认的分类法，超出了我们对线性模型所下的狭义定义6.1，它的依据是区分参数 $\beta_1, \cdots, \beta_p$ 的性质．从某种意义上说 β_j 反映了自变量的第 j 个因子对观察值 y 的影响的大小，常称之为第 j 个**因子的效应**．这个效应在实际问题里可以是随机的．于是，可依效应的随机与否来分类：如果 $\beta_1, \cdots, \beta_p$ 都不是随机变量而是固定的未知参数，则称模型为**固定效应模型**；如果 $\beta_1, \cdots, \beta_p$ 都是不可观察的随机变量，就称模型为**随机效应模型**；当 $\beta_1, \cdots, \beta_p$ 中既有未知参数又有不可观察的随机变量，就称它**混合效应模型**．本书只限于讨论固定效应模型．随机效应情形的讨论可参看有关线性模型的专著．

就我们将要讨论的三种统计模型来说，所侧重的统计问题是有区别的．回归分析模型大部分来源于非人力所能控制的观测，它的主要统计问题是根据已得的数据给出回归方程，其主要目的是用来预测(或称预报)．方差分析模型大都来源于试验，一般是在有了精心的设计之后，它的主要统计问题是根据数据去检验因

子效应的显著性(即对观察值的影响是否显著)，其主要目标是寻求最佳试验点．协方差分析则介乎其中．然而，根据不同统计问题所发展的方法，往往并不局限于原模型．就不同类的模型而言，往往也有本质上完全相同的统计问题．即使上面各类中提出的主要问题，也可从理论上统一处理．下面我们先分节讨论针对各类模型所发展的统计分析方法，然后在第5节中就一般模型的理论问题作一些探讨．

§6.2 回归分析

(一)回归的概念

回归分析的目的是寻求一个随机变量[1] Y 对一组随机变量 $X_1, \cdots, X_p$(当 $p=1$，就是一个随机变量)的统计依赖关系．统计依赖关系不再是单纯的因果关系，它与一般变量间的函数关系有本质不同．但这种依赖关系是在一定的统计意义下确实存在的，我们将逐步明确它的含义．

回归(regression)这一术语是1886年Galton在研究遗传现象时引进的．他发现：虽然高个子的先代会有高个子的后代，但后代的增高并不与先代的增高等量．他称这一现象为“向平常高度的回归”．尔后，他的朋友K. Pearson等人搜集了上千个家庭成员的身高数据，分析出儿子的身高 y 和父亲的身高 x 大致可归结为以下关系

$$y=0.516x+33.73. \text{（以英寸为单位）}$$

由于 $0.516\doteq 0.5$，意味着如父亲身高超过父亲平均身高6英寸，则其儿子的身高，大约只超过儿子平均身高3英寸，可见有向平均值返回之趋势．诚然如今对回归这一概念的理解并不是Galton的原意，但这一名词却一直沿用下来，成为统计中最常用的概念之一．

让我们先从观察值出发来讨论．设在一个总体中取得某个样

1) 在多元统计分析中，将把这里的一个随机变量推广为一组随机变量．

本，观察它的两个特征，得到反映这两个特征的指标(X, Y)的观察值(x_1, y_1)，将这样的观察进行n次，获得观察值(x_α, y_α)，$\alpha=1, \cdots, n$. 从而得到平面上n个点，如图6.1. 在n较大的情况下(n太小就不足为凭)，如果有一条曲线基本上通过这些点，或使这些点的大部分偏离曲线不远，则称这条曲线是对观察值的拟合(曲线). 如果这条曲线的方程能表成$y=f(x)$，则称此曲线为y对x的回归曲线. 当此曲线是直线时，就称之为回归直线. 这暂且还是一个很粗糙的想法，但却是回归概念的一个直观的出发点.

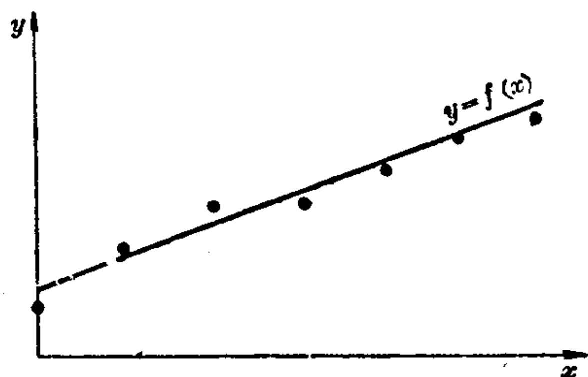


图 6.1

现在我们从理论的角度予以严格讨论. 随后再回到对观察值的回归分析上来. 设(X, Y)有联合概率分布，并且存在二阶矩，那么，当X取某个特殊值x时，Y虽不能完全确定，却有一个确定的条件分布$P(\cdot|x)$，从而有一个确定的条件均值$E(Y|x)$是x的函数. 将函数$y=E(Y|x)$的图象表示在图6.2中. 图中曲线(虚线的)是Y的条件密度的示意图，意味着Y按此分布散布在曲线$y=E(Y|x)$的上下. 从而有

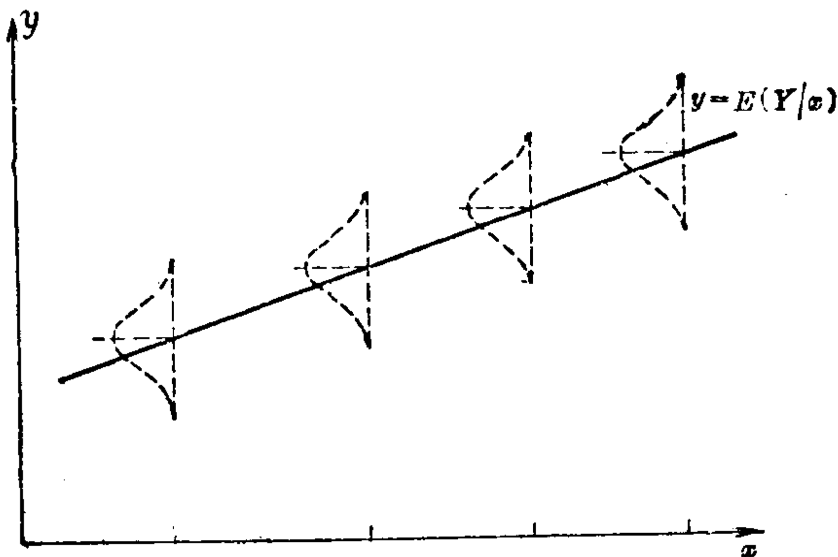


图 6.2

定义 6.2 设(X, Y)有联合概率分布，如果对 $\boldsymbol{X}$ 的每一给定值 $\boldsymbol{x}$，都存在 Y 的条件期望，则称此条件期望 $E(Y|\boldsymbol{x})$为 Y 对 $\boldsymbol{x}$ 的**回归函数**，简称为**回归**. 当 $\boldsymbol{X}$ 是一维的，称作一元回归，当 $\boldsymbol{X}$ 是 p 维向量，就称为 p 元回归. $y=E(Y|\boldsymbol{x})$的图象就称为**回归曲线**(当 $\boldsymbol{X}$ 是一维的)或**回归曲面**(当 $\boldsymbol{X}$ 是多维的). 为了区别于后面将要引进的经验回归，称这里定义的回归为**理论回归**或**总体回归**. 但不致混淆时不予区分.

回归中的自变量 $\boldsymbol{X}$ 常称为**回归因子**或**预报因子**. 回归问题中的 Y 称为因变量，**回归量**或**预报量**、Y 在试验点 $\boldsymbol{x}$ 的值是随机变量，为说明与 x 有关可记为 Y_x. $E(Y|\boldsymbol{x})$被看作对 Y_x 的一个预报，有时用记号 $\hat{Y}_x=E(Y|\boldsymbol{x})$示之.

在回归问题中，视 Y 和 $\boldsymbol{X}$ 的地位是不对称的. 但在许多情形下，这可能是人为的处理. 例如我们可考虑体重对身高的回归，也可考虑身高对体重的回归. 不过要注意这样所得的两条回归曲线，一般并不重合. 可参看习题 6.7. 当然 Y 和 $\boldsymbol{X}$ 地位不对称的情形也很常见，譬如我们不能作父亲身高对儿子身高的回归，因为说父亲身高统计依赖于儿子身高显然是荒谬的. 所以，在实际问题中不能乱用回归方法.

现在我们面临一个必须回答的理论问题：将 $\boldsymbol{x}$ 的其它函数，如 $f(\boldsymbol{x})$，看作 Y 对 $\boldsymbol{X}$ 的"回归"(理解为对 Y_x 的预报)是否有可能比 $E(Y|\boldsymbol{x})$更好？要回答这个问题，先得确定一个好坏的标准. 这当然要以 Y_x 与 $f(\boldsymbol{x})$的差距来衡量. 注意到样本的二重性(它也是随机变量)，我们可以用均方误差 $E[Y-f(\boldsymbol{x})]^2$ 作为好坏的标准.

定义 6.3 设$(Y, \boldsymbol{X})$有联合分布且存在二阶矩. 如果 $M(\boldsymbol{X})$满足

$$E[Y-M(\boldsymbol{X})]^2=\min_f E[Y-f(\boldsymbol{X})]^2, \tag{6.10}$$

则称 $M(\boldsymbol{X})$为对 Y 的**最小均方误差预测**.

我们有

定理 6.1 在定义 6.3 的条件下，$E(Y|\boldsymbol{X})$是对 Y 的最小均

方误差预测．并且，在 Y 的一切预测中，$E(Y|\boldsymbol{X})$ 与 Y 的相关系数达到极大值，这个极大值是

$$\sqrt{D(M)}/\sqrt{D(Y)}, \tag{6.11}$$

其中 $D(M)$ 是 $M(\boldsymbol{X})\triangleq E(Y|\boldsymbol{X})$ 的方差.

证 对一切(可测)函数 $f(\boldsymbol{X})$，我们有

$$\begin{aligned}E[Y-f(\boldsymbol{X})]^2&=E[Y-E(Y|\boldsymbol{X})+E(Y|\boldsymbol{X})-f(\boldsymbol{X})]^2\\&=E[Y-E(Y|\boldsymbol{X})]^2+E[E(Y|\boldsymbol{X})\\&\quad-f(\boldsymbol{X})]^2+2E[Y-E(Y|\boldsymbol{X})][E(Y|\boldsymbol{X})\\&\quad-f(\boldsymbol{X})].\end{aligned}$$

由条件期望的基本性质可得

$$\begin{aligned}&E[Y-E(Y|\boldsymbol{X})][E(Y|\boldsymbol{X})-f(\boldsymbol{X})]\\&=EE\{[Y-E(Y|\boldsymbol{X})][E(Y|\boldsymbol{X})-f(\boldsymbol{X})]|\boldsymbol{X}\}\\&=E\{[E(Y|\boldsymbol{X})-f(\boldsymbol{X})]E\{[Y-E(Y|\boldsymbol{X})]|\boldsymbol{X}\}\}\\&=0.\end{aligned}$$

从而得 $E[Y-f(\boldsymbol{X})]^2\geqslant E[Y-E(Y|\boldsymbol{X})]^2$. 得 $E(Y|\boldsymbol{X})\triangleq M(\boldsymbol{X})$ 是对 Y 的最小均方误差预测.

容易计算

$$\begin{aligned}\mathrm{Cov}(Y,f)&=E(Y-EY)[f(\boldsymbol{X})-Ef(\boldsymbol{X})]\\&=EE[(Y-EY)[f(\boldsymbol{X})-Ef(\boldsymbol{X})]|\boldsymbol{X}]\\&=E[f(\boldsymbol{X})-Ef(\boldsymbol{X})][E(Y|\boldsymbol{X})-EY]\\&=\mathrm{Cov}(f,M).\end{aligned}$$

将上式中 f 用 M 代入可得 $\mathrm{Cov}(Y,M)=D(M)$. 根据 Schwartz 不等式推出相关系数间的关系式如下

$$\begin{aligned}\rho(Y,f)&=\frac{\mathrm{Cov}(Y,f)}{\sqrt{D(Y)D(f)}}=\frac{\mathrm{Cov}(M,f)}{\sqrt{D(Y)D(f)}}\\&\leqslant\frac{\sqrt{D(M)D(f)}}{\sqrt{D(Y)D(f)}}=\frac{D(M)}{\sqrt{D(Y)D(M)}}\\&=\rho(Y,M).\end{aligned}$$

定理至此证毕.

由定理 6.1 的证明可以看出，用 $E(Y|\boldsymbol{x})$ 来预测 Y，其预测误差 $\varepsilon \triangleq Y-E(Y|\boldsymbol{x})$ 满足

$$E\varepsilon=0,\ D(\varepsilon)=E\varepsilon^2=D(Y)-D(M).$$

引进**预测精度** $\lambda \triangleq D(\varepsilon)/D(Y)$，有

$$\lambda=1-D(M)/D(Y)=1-\rho^2(Y,\ M). \tag{6.12}$$

可见预测精度是一个与相关系数 $\rho(Y,\ M)$ 紧密相依的量．在统计中称 Y 和 $\boldsymbol{X}$ 的线性函数的相关系数的极大值 $\rho_{Y,X}$ 为 Y 和 $\boldsymbol{X}$ 的**多重相关系数**．定理 6.1 表示 $\rho_{Y,X}\leqslant\rho(Y,\ M)$．

在 $\boldsymbol{X}$ 给定 $\boldsymbol{x}$ 的条件下，记相应的 Y 为 $Y_{\boldsymbol{x}}$，令 $\varepsilon_{\boldsymbol{x}} \triangleq Y_{\boldsymbol{x}}-E(Y|\boldsymbol{x})$，有 $E(\varepsilon_{\boldsymbol{x}}|\boldsymbol{x})=0$．于是有

$$Y_{\boldsymbol{x}}=E(Y|\boldsymbol{x})+\varepsilon_{\boldsymbol{x}}. \tag{6.13}$$

称(6.13)为**回归模型**．其中 $E(Y|\boldsymbol{x})$ 称为回归模型的**主要部分**，或如前面已经定义的叫作 Y 对 $\boldsymbol{x}$ 的回归函数．

当回归 $E(Y|\boldsymbol{x})$ 是 $\boldsymbol{x}$ 的线性函数，即

$$E(Y|\boldsymbol{x})=\beta_0+\beta_1x_1+\cdots+\beta_px_p,$$

则回归模型(6.13)可记为

$$Y_{\boldsymbol{x}}=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon_{\boldsymbol{x}}, \tag{6.14}$$

称作**线性回归模型**[1]．为了记号上的方便，我们有时省略(6.14)中作为角标的 x，但此时不能再把 $x_1,\ \cdots,\ x_p$ 依样本二重性理解为随机变量，我们已把它作为一般变量对待．在这一理解下，(6.14)如定义 6.1 所述是一个线性模型．以后就将(6.14)记为

$$Y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon. \tag{6.14$'$}$$

然而在实际问题中，$E(Y|\boldsymbol{x})$ 不一定是 $\boldsymbol{x}$ 的线性函数，这时如人为地用 $\boldsymbol{x}$ 的线性函数去预测 Y，有可能使预测精度变坏．但用线性函数作预测自有它简便之处，况且下面关于正态总体的一个结果表明，在正态情形下，限于在线性函数类中选取最佳预测，不会有任何损失．而正态情形无疑是最重要的情形．

---

1) 我们在定义 6.1 中给出的线性模型的定义是指对未知参数线性．这里从回归的角度强调了对自变量的线性性，切勿与原概念混淆．就回归而言非线性的，如多项式回归，仍属线性模型．看习题 1，2．

设 $\boldsymbol{z}=(y, x_1, \cdots, x_p)'$ 遵从 $p+1$ 维正态分布，有密度函数为

$$f(\boldsymbol{z})=(2\pi)^{-\frac{p+1}{2}}(\det \boldsymbol{V})^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\nu})'\boldsymbol{V}^{-1}(\boldsymbol{z}-\boldsymbol{\nu})\right\}.$$

记
$$\boldsymbol{V}^{-1}=\begin{pmatrix} v^{11} & \boldsymbol{\nu}' \\ \boldsymbol{\nu} & \boldsymbol{V}^{22}\end{pmatrix}\begin{matrix}1\\p\end{matrix}, \quad Ey=\theta, \ E\boldsymbol{x}=\boldsymbol{\mu},$$

则记 $\boldsymbol{x}$ 的边缘密度为 $\varphi(\boldsymbol{x})$ 就有

$$E(Y|\boldsymbol{x})=\int_{-\infty}^{\infty} y(2\pi)^{-\frac{p+1}{2}}(\det \boldsymbol{V})^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\nu})'\boldsymbol{V}^{-1}(\boldsymbol{z}-\boldsymbol{\nu})\right\}\varphi(\boldsymbol{x})^{-1}dy$$

将 $Q\triangleq(\boldsymbol{z}-\boldsymbol{\nu})'\boldsymbol{V}^{-1}(\boldsymbol{z}-\boldsymbol{\nu})$ 依 y 重新配方得

$$\begin{aligned} Q &=(y-\theta)^2v^{11}+2(y-\theta)\boldsymbol{\nu}'(\boldsymbol{x}-\boldsymbol{\mu})+(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{V}^{22}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &=v^{11}[y-\theta+(v^{11})^{-1}\boldsymbol{\nu}'(\boldsymbol{x}-\boldsymbol{\mu})]^2+c \quad (c\text{ 不依赖 }y). \end{aligned}$$

从而可知 Y 的条件密度是正态的，故可得

$$E(Y|\boldsymbol{x})=\theta-(v^{11})^{-1}\boldsymbol{\nu}'(\boldsymbol{x}-\boldsymbol{\mu}) \tag{6.15}$$

是 $\boldsymbol{x}$ 的线性函数.

定义 6.4 设 $(Y, \boldsymbol{X})$ 有联合分布，且存在二阶矩. 如果在 $\boldsymbol{X}$ 的线性函数类 $\{c_0+\mathbf{c}'\boldsymbol{x}\}$ 中 $M(\boldsymbol{x})$ 满足

$$E[Y-M(\boldsymbol{x})]^2=\min_{c_0,c} E[Y-c_0+\mathbf{c}'\boldsymbol{x}]^2 \tag{6.16}$$

则称 $M(\boldsymbol{x})$ 是对 Y 的最小均方误差线性预测.

在已知 $(Y, \boldsymbol{x})$ 的二阶矩时，$M(\boldsymbol{x})$ 不难求出. 我们有

定理 6.2

设 $(Y, \boldsymbol{X})$ 有联合分布，其协方差阵

$$\operatorname{Cov}\begin{pmatrix} Y \\ \boldsymbol{X}\end{pmatrix}\triangleq\begin{pmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{XY} & \Sigma_{XX}\end{pmatrix}, \text{ 其中 } \Sigma_{XX} \text{ 可逆}, \tag{6.17}$$

这里 $\sigma_Y^2=D(Y)$, $\sigma_{YX}=\sigma_{XY}'=(\operatorname{Cov}(Y, X_1), \cdots, \operatorname{Cov}(Y, X_p))$, $\Sigma_{XX}=\operatorname{Cov}\boldsymbol{X}$. 则有 Y 的最小均方误差线性预测为

$$M(\boldsymbol{x})=EY-\sigma_{YX}\Sigma_{XX}^{-1}E\boldsymbol{X}+\sigma_{YX}\Sigma_{XX}^{-1}\boldsymbol{x} \tag{6.18}$$

证 设 $\boldsymbol{x}$ 的任一线性函数为 $c_0+\mathbf{c}'\boldsymbol{x}$. 我们有

$$\begin{aligned} E(Y-c_0-\mathbf{c}'\boldsymbol{x})^2 &=E[(Y-EY)-\mathbf{c}'(\boldsymbol{x}-E\boldsymbol{x}) \\ &\quad +(EY-c_0-\mathbf{c}'E\boldsymbol{x})]^2 \end{aligned}$$

$$
\begin{aligned}
&=\sigma_Y^2+\mathbf{c}'\Sigma_{XX}\mathbf{c}-2\sigma_{YX}\mathbf{c}+(EY-c_0-\mathbf{c}'E\boldsymbol{x})^2\\
&=\sigma_Y^2+(\Sigma_{XX}^{1/2}\mathbf{c}-\Sigma_{XX}^{-1/2}\sigma_{XY})'(\Sigma_{XX}^{1}\mathbf{c}-\Sigma_{XX}^{-1/2}\sigma_{XY})\\
&\quad-\sigma_{YX}\Sigma_{XX}^{-1}\sigma_{XY}+(EY-c_0-\mathbf{c}'E\boldsymbol{x})^2.
\end{aligned}
$$

从而极小值点 c_0, $\mathbf{c}$ 是方程

$$
\begin{cases}
\Sigma_{XX}^{1/2}c-\Sigma_{XX}^{-1/2}\sigma_{XY}=0\\
EY-c_0-\mathbf{c}'E\boldsymbol{x}=0
\end{cases}
$$

的解. 易见解为

$$
\mathbf{c}=\Sigma_{XX}^{-1}\sigma_{XY},\ c_0=EY-\sigma_{YX}\Sigma_{XX}^{-1}E\boldsymbol{x}.\ \text{证毕.}
$$

由附录 A. 5.2, 不难验证在联合分布为正态时, 由(6.18)给出的 $M(\boldsymbol{x})$与(6.15)中的 $E(Y|\boldsymbol{x})$一致. 从另一角度看, 这种一致性当然已经由定理 6.1 和(6.15)式所保证.

我们称(6.18)所给出的 $M(\boldsymbol{x})$为 Y **对 $\boldsymbol{x}$ 的线性回归**. 容易计算它对 Y 的预测也是无偏的,其预测精度为

$$
\lambda_L=1-\sigma_{YX}\Sigma_{XX}^{-1}\sigma_{XY}/\sigma_Y^2=1-\rho_{Y,X}^2 \tag{6.19}
$$

λ_L 一般将大于(6.12)中的 λ.

记 $\beta_0=EY-\sigma_{YX}\Sigma_{xx}^{-1}E_X$, $\beta=\Sigma_{xx}^{-1}\sigma_{XY}$, 得**总体线性回归模型**为

$$
Y=\beta_0+\boldsymbol{\beta}'\boldsymbol{x}+\varepsilon \tag{6.20}
$$

称

$$
y=\beta_0+\boldsymbol{\beta}'\boldsymbol{x} \tag{6.21}
$$

为**回归方程**. (注意这里 y 是函数值的记号,不要混淆为 Y 的观察值.) $\boldsymbol{\beta}$ **称作回归系数**.

在整个这一段中, 我们并未限制自变量 $\boldsymbol{x}$ 的性质, 因此这里所讲的回归模型是一般的, 它与第一节所定义的回归分析模型略有差别.

(二)回归系数的估计与经验回归

在总体线性回归方程中,回归系数是由总体的二阶矩给定的. 但在实际问题中,二阶矩一般是未知的,要用取自总体的样本来估计.

设容量为 n 的简单随机样本记为

$$\boldsymbol{y}=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\quad \boldsymbol{X}=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix},\quad \text{且记 } \mathbf{1}=\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

得样本模型为

$$\boldsymbol{y}=(\mathbf{1}\boldsymbol{X})\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix}+\boldsymbol{\varepsilon} \quad (\text{在 } X \text{ 给定的条件下}). \qquad (6.22)$$

现在我们假定(6.22)是一个回归分析模型. 这时可认为设计矩阵的秩 $rk(\mathbf{1X})=p+1$. 这是因为 $\boldsymbol{X}$ 的元素是容许取连续变化的值的, 降秩的情形很少出现, 在理论推导中就不顾及这种例外了. 如果在实际工作中遇到降秩情形, 可另予适当处理.

为了符号简单, 我们改记

$$\boldsymbol{Z}=(\mathbf{1X}),\ \boldsymbol{\theta}=(\beta_0\boldsymbol{\beta}')',$$

从而将(6.22)改记为

$$\boldsymbol{y}=\boldsymbol{Z\theta}+\boldsymbol{\varepsilon} \qquad (6.22')$$

用最小二乘法估计 $\boldsymbol{\theta}$, 即求 $\boldsymbol{\theta}$ 的估计量 $\hat{\theta}$ 使满足

$$\|\boldsymbol{y}-\boldsymbol{z}\hat{\boldsymbol{\theta}}\|^2=\min_{\theta\in R^{p+1}}\|\boldsymbol{y}-\boldsymbol{z}\,\boldsymbol{\theta}\|^2 \qquad (6.23)$$

其中 $\|\boldsymbol{a}\|^2=\boldsymbol{a}'\boldsymbol{a}=\Sigma_1^n a_i^2$ 表示 $\boldsymbol{a}$ 的欧氏范数的平方. 用数学分析的求极值法当然不难给出(6.23)的解. 但我们宁可用代数方法去求解, 这一方法称作**平方和分解法**, 它在以后还会经常用到.

假定 $\hat{\boldsymbol{\theta}}$ 是(6.23)的解, 而 $\boldsymbol{\theta}$ 是任意的, 我们有

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{z\theta}\|^2&=\|\boldsymbol{y}-\boldsymbol{z}\hat{\boldsymbol{\theta}}+\boldsymbol{z}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|^2\\&=\|\boldsymbol{y}-\boldsymbol{z}\hat{\boldsymbol{\theta}}\|^2+\|\boldsymbol{z}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|^2+2(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\boldsymbol{z}'(\boldsymbol{y}-\boldsymbol{z}\hat{\boldsymbol{\theta}}).\end{aligned}$$

于是, 欲使

$$\|\boldsymbol{y}-\boldsymbol{z\theta}\|^2\geqslant\|\boldsymbol{y}-\boldsymbol{z}\hat{\boldsymbol{\theta}}\|^2 \quad \text{对一切 } \boldsymbol{\theta}\in R^{p+1} \text{ 成立}$$

的充要条件是

$$(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\boldsymbol{z}'(\boldsymbol{y}-\boldsymbol{z}\hat{\boldsymbol{\theta}})=0 \quad \text{对一切 } \boldsymbol{\theta}\in R^{p+1} \text{ 成立}.$$

上式对一切 $\boldsymbol{\theta}$ 成立又等价于 $\boldsymbol{z}'(\boldsymbol{y}-\boldsymbol{z}\hat{\boldsymbol{\theta}})=0$, 即

$$\boldsymbol{z}'\boldsymbol{z}\hat{\boldsymbol{\theta}}=\boldsymbol{z}'\boldsymbol{y} \qquad (6.24)$$

称(6.24)为模型(6.22′)的正规方程. 它的解

$$\hat{\boldsymbol{\theta}}=(\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{y} \tag{6.25}$$

就是 $\boldsymbol{\theta}$ 的最小二乘估计．记 $\hat{\boldsymbol{\theta}}=(\hat{\beta}_0\hat{\boldsymbol{\beta}}')'$，就得到 β_0 和 $\boldsymbol{\beta}$ 的估计，称之为经验回归系数．相应给出的方程

$$\hat{y}=\hat{\beta}_0+\hat{\boldsymbol{\beta}}'\boldsymbol{x} \tag{6.26}$$

称作 Y 对 $\boldsymbol{x}$ 的经验回归方程．称 $\hat{\beta}_0+\hat{\boldsymbol{\beta}}'\boldsymbol{x}$ 为经验回归函数．

用原来的数据来表出 $\hat{\beta}_0$ 和 $\hat{\boldsymbol{\beta}}$．在 $p=1$ 的情形下计算较简单．此时

$$(\boldsymbol{z}'\boldsymbol{z})^{-1}=\left[\begin{pmatrix}\mathbf{1}'\\ \boldsymbol{x}'\end{pmatrix}(\mathbf{1}\boldsymbol{x})\right]^{-1}=\begin{pmatrix}n & \Sigma x_\alpha\\ \Sigma x_\alpha & \Sigma x_\alpha^2\end{pmatrix}^{-1}$$

$$=(n\Sigma x_\alpha^2-(\Sigma x_\alpha)^2)^{-1}\begin{pmatrix}\Sigma x_\alpha^2 & -\Sigma x_\alpha\\ -\Sigma x_\alpha & n\end{pmatrix},$$

$$\boldsymbol{z}'\boldsymbol{y}=\begin{pmatrix}\Sigma y_\alpha\\ \Sigma x_\alpha y_\alpha\end{pmatrix},$$

得

$$\begin{aligned}\hat{\beta}_0&=(\Sigma x_\alpha^2\Sigma y_\alpha-\Sigma x_\alpha\Sigma x_\alpha y_\alpha)/[n\Sigma x_\alpha^2-(\Sigma x_\alpha)^2],\\ \hat{\boldsymbol{\beta}}_1&=(n\Sigma x_\alpha y_\alpha+\Sigma x_\alpha\Sigma y_\alpha)/[n\Sigma x_\alpha^2-(\Sigma x_\alpha)^2].\end{aligned} \tag{6.27}$$

在一般情形下，有

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\boldsymbol{\beta}}\end{pmatrix}=\begin{pmatrix}n & \mathbf{1}'\boldsymbol{X}\\ \boldsymbol{X}'\mathbf{1} & \boldsymbol{X}'\boldsymbol{X}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{1}'\boldsymbol{y}\\ \boldsymbol{X}'\boldsymbol{y}\end{pmatrix}. \tag{6.28}$$

具体计算较繁时，可借助计算机去完成．

在 $p=1$ 时(**6.27**)的结果，与在总体线性回归(6.18)中用样本矩去代替总体矩所得结果完全一致．其实，对于一般的 p，此结论也对，将在第七章讨论回归时论及．

经验回归方程(6.26)虽然得来容易，却不可轻视它的意义．可以说它概括了历史资料所提供的 Y 对 $\boldsymbol{X}$ 的线性统计依赖的信息(当然是在 $E(Y|\boldsymbol{X})$ 为线性或接近线性的前提下)，因此，用 $\hat{y}_*=\hat{\beta}_0+\hat{\boldsymbol{\beta}}'\boldsymbol{x}_*$ 来预测 Y 在新的试验点 $\boldsymbol{x}_*$ 上的观察值 y_*，似不会有过大偏差．这种预测，在研讨未来对现在的依赖关系时十分必要，如气象、地震、洪水的预报以及社会、经济趋势的预测等；在研究对象中途丢失或数据缺损的情况下，y_* 实际上变得不可观察，预测当然

不可缺少；在试验或观测需要耗费过大时，预测也很有价值．然而，在用回归方程作预测时，务必相当谨慎．因为模型是否确系线性，因子是否选择适当，数据是否充足和可靠，都是值得怀疑的．要取得良好的预测效果，还需依赖对于实际问题的具体知识．

为了符号上的简便，我们可以对线性回归模型作合理的简化．一种办法是引入假变量 x_0，令其在各次试验中均取 1，从而不认为回归模型有特殊的常数项，并不妨把 $\boldsymbol{X}$ 仍视为 p 列，有

$$\boldsymbol{y}=\boldsymbol{x}\boldsymbol{\beta}+\varepsilon \tag{6.29}$$

另一办法是作数据变换，因回归中的**总平均** $\hat{\beta}_0$ 不过反映数据计算的起始值．如对数据作变换：将每个数据减去它所在列的数据平均值，即 $y_a\mapsto y_a-\frac{1}{n}\sum_1^n y_a$，$x_{aj}\mapsto x_{aj}-\frac{1}{n}\sum_1^n x_{aj}$，$j=1, \cdots, p$，就以每列的平均值作为新的计算起点．这样就可使各变量的样本均值为零，从而不难验证 $\hat{\beta}_0=0$．只需注意在用新的回归方程 $y=\hat{\boldsymbol{\beta}}'\boldsymbol{x}$ 作预测时，y、$\boldsymbol{x}$ 的值都以上述平均值为起始点，则对问题无任何实质性影响．所以，在以后各段的讨论中，除非另有说明的特殊情况，均略去常数项．

（三）预测区域

上段所讲到的预测是没有概率保证的，即我们无法回答：预测值和它要预测的对象吻合的概率是多少？这种情形有点象点估计．现在我们希望能用样本确定一个区域，使得它包含待预测的 y_* 的概率不低于某个被称为置信系数的值 $1-\alpha$．这里 α 是一个根据需要选择的接近于零的正数（如 0.01, 0.05 等）．这就是**预测区域**问题．

在讨论预测区域时，我们必须对总体的分布形式作出假定，否则将无从计算概率．由于在回归模型中我们仅涉及 $\boldsymbol{X}$ 给定时 Y 的条件分布，故把 $\boldsymbol{X}$ 看作一般变量，谈到 Y 的分布时也不再提条件二字．我们仅就 Y 服从正态分布的情形进行讨论，且设各次试验的观察值是互不相关和同方差 σ^2 的，记作 $\boldsymbol{y}\sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$．

首先在上述分布假定下给出回归中涉及的并将在预测区域和下段显著性检验中用到的一些统计量的分布.

在第一章§1.4抽样分布的第(三)段中，我们已经给出本章要用的若干分布及其基本性质,这里还需要作些补充.

引理 6.1

设 $\boldsymbol{Y}\sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$, $\boldsymbol{A}$ 是 n 阶对称阵, $\boldsymbol{B}$ 是 $n\times m$ 阶阵. 则有

(i) $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}\sim\chi_r^2(\delta)\Leftrightarrow\boldsymbol{A}$ 是对称幂等阵，其中 $r=rk\boldsymbol{A}$, $\delta^2=\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$.

(ii) 当 $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}\sim\chi_r^2(\delta)$,

$$\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}\text{与}\boldsymbol{B}'\boldsymbol{Y}\text{独立}\Leftrightarrow\boldsymbol{B}'\boldsymbol{A}=0.$$

证 我们只证两个命题的充分性，而把必要性留作习题. 由于 $\boldsymbol{A}$ 是对称幂等阵, $\boldsymbol{I}-\boldsymbol{A}$ 亦是, 故 $\boldsymbol{A}+(\boldsymbol{I}-\boldsymbol{A})=\boldsymbol{I}$, 且 $rk\boldsymbol{A}+rk(\boldsymbol{I}-\boldsymbol{A})=n$. 故由 Cochran 定理[1](定理 1.2)得 $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}\sim\chi_r^2(\delta)$. (i)的充分性得证. 因(ii)中设 $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}\sim\chi_r^2(\delta)$, 知 $\boldsymbol{A}$ 为对称幂等阵,存在正交阵 $\boldsymbol{U}$ 使 $\boldsymbol{U}'\boldsymbol{A}\boldsymbol{U}=\begin{pmatrix}\boldsymbol{I}_r & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}$. 记 $\boldsymbol{U}=(\boldsymbol{U}_1\ \vdots\ \boldsymbol{U}_2)$, $\boldsymbol{U}_1$ 是 $n\times r$ 阶阵. 令 $\boldsymbol{Y}=\boldsymbol{U}\boldsymbol{X}$,则有

$$\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}=\boldsymbol{X}'\boldsymbol{U}'\boldsymbol{A}\boldsymbol{U}\boldsymbol{X}=\Sigma_1^r x_i^2,\ \boldsymbol{B}'\boldsymbol{Y}=\boldsymbol{B}'\boldsymbol{U}\boldsymbol{X}.$$

注意到 $\boldsymbol{B}'\boldsymbol{A}=\boldsymbol{B}'\boldsymbol{U}\begin{pmatrix}\boldsymbol{I}_r & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{U}'=0$, 有 $\boldsymbol{B}'\boldsymbol{U}_1=0$. 故得 $\boldsymbol{B}'\boldsymbol{Y}=(\boldsymbol{B}'\boldsymbol{U}_1\ \vdots\ \boldsymbol{B}'\boldsymbol{U}_2)\boldsymbol{X}=(\boldsymbol{0}\ \vdots\ \boldsymbol{B}'\boldsymbol{U}_2)\boldsymbol{X}=\boldsymbol{B}'\boldsymbol{U}_2\boldsymbol{X}^{(2)}$, 这里 $\boldsymbol{X}^{(2)}=(X_{r+1}, \cdots, X_n)'$. 由 $\boldsymbol{X}$ 分量的相互独立性,得 $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$ 与 $\boldsymbol{B}'\boldsymbol{Y}$ 独立.

记 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是回归模型(6.29)中 $\boldsymbol{\beta}$ 的最小二乘估计,称 $\|\boldsymbol{y}-\boldsymbol{x}\hat{\boldsymbol{\beta}}\|^2$ 为**剩余平方和**(或称**残差平方和**),记为 S_ε^2. 易见

$$S_\varepsilon^2=\boldsymbol{y}'\boldsymbol{P}_{x^\perp}\boldsymbol{y},\ \boldsymbol{P}_{x^\perp}=\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{ 是正投影阵.}$$

而 $\hat{y}\triangleq\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{P}_x\boldsymbol{y}$, 这里 $\boldsymbol{P}_x=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 是到 $\mu(\boldsymbol{X})$($\boldsymbol{X}$ 的列向量张成的线性空间)的正投影阵. 于是有

定理 6.3

在上述记号下有

---

1) Cochrar 定理的证明蕴涵了对非中心情形的结论。

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),\ \hat{\boldsymbol{y}} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{P}_x),\ \sigma^{-2}S_\varepsilon^2 \sim \chi^2_{n-p}, \tag{6.30}$$

且 $\hat{\boldsymbol{\beta}}$、$\hat{\boldsymbol{y}}$ 和 S_ε^2 独立．其中 p 是设计阵 $\boldsymbol{X}$ 的列数．

证 由引理6.1，只需注意到 $(\boldsymbol{X\beta})'\boldsymbol{P}_x\boldsymbol{X\beta} = 0$，且 $rk\boldsymbol{P}_x = t_r\boldsymbol{P}_x = n - t_r\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = n - t_r(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = n - p$,立得结论．需要指出的是这里 $\hat{\boldsymbol{y}}$ 的协方差矩阵是退化的，这种情形下多元正态分布的确切含义将在第七章 §7.1 中给出，故这里的结论暂时还有一定模糊性．

设在历史样本的基础上已得回归方程

$$\hat{y} = \boldsymbol{x}'\hat{\boldsymbol{\beta}}$$

设新的试验点是 $\boldsymbol{x}_* = (x_1, \cdots, x_p)'$．为求 $\boldsymbol{x}_*$ 上观察值 y_* 的预测区域,需要寻找有已知分布的适当的统计量(它应当含有 y_*)．因 $y_* \sim N_1(\boldsymbol{x}_*'\boldsymbol{\beta}, \sigma^2)$ 令 $\hat{y}_* = \boldsymbol{x}_*'\hat{\boldsymbol{\beta}} = \boldsymbol{x}_*'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$,得

$$y_* - \hat{y}_* \sim N_1(0, \sigma_*^2),$$

因 y_* 与 $\boldsymbol{y}$ 独立,因而也与 $\hat{y}_*$ 独立,故其中

$$\sigma_*^2 = \sigma^2 + \sigma^2\boldsymbol{x}_*'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_* \triangleq \sigma^2\rho^2,\ \rho^2 = 1 + \boldsymbol{x}_*'(\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}_*.$$

由定理6.3知 $\sigma^{-2}S_\varepsilon^2 \sim \chi^2_{n-p}$,且 S_ε^2 与 $y_* - \hat{y}_*$ 独立．故有

$$T \triangleq \frac{y_* - \hat{y}_*}{S_\varepsilon\rho}\sqrt{n-p} \sim t_{n-p} \quad (t\text{-分布}). \tag{6.31}$$

记 $t_{n-p}(\alpha)$是分布 t_{n-p} 的上侧 α-分位点，即 $t_{n-p}(\alpha)$满足 $P(T \geqslant t_{n-p}(\alpha)) = \alpha$．现设置信系数为 $1-\alpha$,则有 $P\left(|T| \leqslant t_{n-p}\left(\frac{\alpha}{2}\right)\right) = 1-\alpha$．其中 $|T| \leqslant t_{n-p}\left(\frac{\alpha}{2}\right)$等价于

$$\hat{y}_* - t_{n-p}\left(\frac{\alpha}{2}\right)S_\varepsilon\rho(n-p)^{-\frac{1}{2}} \leqslant y_* \leqslant \hat{y}_* + t_{n-p}\left(\frac{\alpha}{2}\right)S_\varepsilon\rho(n-p)^{-\frac{1}{2}}.$$

记上面不等式的左右两端分别为 y_1 和 y_2，遂得区间$[y_1, y_2]$包含 y_* 的概率 $P(y_* \in [y_1, y_2]) = 1-\alpha$，因此，$[y_1, y_2]$就是 y_* 的置信系数为 $1-\alpha$ 的预测区间．

如果视 y_1, y_2 为 $\boldsymbol{x}_* = (x_1^*, \cdots, x_p^*)'$ 的函数,那么 $y = y_1$, $y = y_2$ 就是 R^{p+1} 中的两个曲面．这两个曲面所夹的区域称为**预测带**．图

6.3 中画出了 x_* 为一维时的预测带.

同时预测在 m 个试验点 $\boldsymbol{x}_*=(x_{aj})$上的观察值 $\boldsymbol{y}_*=(y_{1*},\cdots,y_{m*})'$ 的预测区域，不难仿照上面的方法的原则给出. 此时 $\hat{\boldsymbol{y}}_*=\boldsymbol{X}_*\hat{\boldsymbol{\beta}}=\boldsymbol{X}_*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, 由于新观察值 $\boldsymbol{y}_*$ 应与原观察值 $\boldsymbol{y}$ 独立，得 $\boldsymbol{y}_*$ 与 $\hat{\boldsymbol{y}}_*$ 独立,有

$$E(\boldsymbol{y}_*-\hat{\boldsymbol{y}}_*)=0,\ \mathrm{Cov}(\boldsymbol{y}_*-\hat{\boldsymbol{y}}_*)=\sigma^2\boldsymbol{I}_m+\sigma^2\boldsymbol{X}_*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_*'.$$

记 $\boldsymbol{\Sigma}=\boldsymbol{I}_m+\boldsymbol{X}_*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_*'$, 它当然是正定阵. 因为 $\sigma^{-1}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{y}_*-\hat{\boldsymbol{y}}_*)\sim N_m(0,\boldsymbol{I}_m)$,有

$$\sigma^{-2}(\boldsymbol{y}^*-\hat{\boldsymbol{y}}_*)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_*-\hat{\boldsymbol{y}}_*)\sim\chi_m^2.$$

因 $\boldsymbol{y}_*$, $\hat{\boldsymbol{y}}_*$ 都与 S_ε^2 独立,我们有

$$F\triangleq\frac{(\boldsymbol{y}_*-\hat{\boldsymbol{y}}_*)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_*-\hat{\boldsymbol{y}}_*)}{S_\varepsilon^2}\cdot\frac{n-p}{m}\sim F_{m,n-p}.$$

记 $F_{m,n-p}(\alpha)$为 $F_{m,n-p}$ 分布的上侧 α-分位点,就有

$$P\left((y_*-\hat{y}_*)'\Sigma^{-1}(y_*-\hat{y}_*)\leqslant\frac{m}{n-p}S_\varepsilon^2F_{m,n-p}(\alpha)\right)=1-\alpha.$$

由于 $c(\boldsymbol{y})\triangleq\left\{\boldsymbol{y}:\ (\boldsymbol{y}-\hat{\boldsymbol{y}}_*)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_*-\hat{\boldsymbol{y}}_*)\leqslant\frac{m}{n-p}S_\varepsilon^2F_{m,n-p}(\alpha)\right\}$是 R^m 中的椭球. 得 $c(\boldsymbol{y})$是 $\boldsymbol{y}_*$ 的置信系数为 $1-\alpha$ 的预测区域(预测椭球).

现在我们讨论预测区域的精度问题. 讨论这一问题的必要性是明显的. 如果预测区域很大,我们说它包含了 y_*,实际上没有什么意义,并不能使我们对 y_* 有比较明确的认识. 我们希望预测区域越小越好. 遗憾的是它受到各方面的制约,不能随心所欲. 在 $m=1$ 时,预测区间的长度是 $\Delta=y_2-y_1$,取它的平方得

$$\Delta^2=(y_2-y_1)^2=\left(t_{n-p}\left(\frac{\alpha}{2}\right)\right)^2s_\varepsilon^2(1+\boldsymbol{x}_*'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_*)(n-p)^{-1}.$$

不难算出它的均值是

$$E\Delta^2=\left(t_{n-p}\left(\frac{\alpha}{2}\right)\right)^2\sigma^2(1+\boldsymbol{x}_*'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_*).$$

它显然大于$\left(t_{n-p}\left(\frac{\alpha}{2}\right)\right)^2\sigma^2$. 尤其当 $\boldsymbol{X}'\boldsymbol{X}$ 有很小的特征值，且 $\boldsymbol{x}_*$

是相应于它的特征向量，$E\Delta^2$ 可以很大．所以，当 $\boldsymbol{X}'\boldsymbol{X}$(称为信息矩阵)有很小的特征值(即它接近于退化)时，预测精度可能极低．这时的信息阵 $\boldsymbol{X}'\boldsymbol{X}$ 称为是病态的，它带来很多麻烦．这种情形下的处置办法已有人作过很多讨论，不在此介绍．

注意到 $E\dfrac{m}{n-p}S_e^2F_{m,n-p}(\alpha)=m\sigma^2F_{m,n-p}(\alpha)$，且 $\boldsymbol{\Sigma}^{-1}$ 的特征值小于1．不难通过较细致的计算验证它的预测精度将低于 $m=1$ 情形．因为要同时预测 m 个，精度的降低是合乎情理的，m 越大精度自然就越低．

最后我们简略地介绍一个与预测相反的问题称作控制．它研究如何选取自变量的变化范围(控制区域)，使得以一定概率保证相应的观察值落在某个给定的区域内．在 $p=1$ 时，

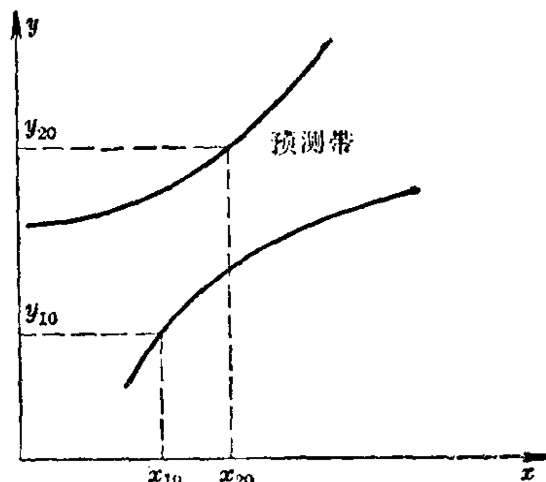


图 6.3

我们可从图6.3中的预测带直观地给出解法．如欲使 y 以 $100(1-\alpha)\%$ 的概率落在区间 (y_{10}, y_{20}) 之内，可取自变量控制区间为 (x_{10}, x_{20})．在此不给出更细致的讨论．

(四)显著性检验

前面我们已经讨论了从样本观察值出发给出线性回归和预测的方法．尽管从正规方程去找出 $\hat{\boldsymbol{\beta}}$ 完全可以在形式上进行，但所得回归函数是否是原模型的主要部分的较好的拟合，却有赖于对模型的基本假定，即它是一个 Y 对 $\boldsymbol{x}$ 的线性回归模型．

要确认模型假定的合理性是一件很不容易的事．这时试验工作者的实际知识往往更重要．过去的经验当然也有助于我们作出抉择．然而，有时候统计工作者的唯一依据就是一批数据．那么，单纯从这批数据出发，有什么积极办法去检验模型假设的合理性呢？一个直观的办法就是去考察回归直线和数据的拟合状况，但这

个方法过于粗疏. 一个有概率依据的办法就是在正态性假定下进行假设检验.

为了要检验模型

$$y=\mathbf{1}\beta_0+X\beta+\varepsilon \tag{6.32}$$

的合理性, 通常提出的一个假设是

$$H_0:\beta=\mathbf{0} \tag{6.33}$$

如果 H_0 被接受, 就有理由认为 Y 对 x 的统计依赖是线性的这一假定很不可信, 从而转向考虑另外的模型. 如果 H_0 被拒绝, 虽然还不能说模型一定是线性的, 但可以认为 Y 与 x 毕竟存在某种程度的相关是可信的, 从而增加了我们对模型的信赖程度. 即使还谈不到有充分的理由, 我们倾向于接受模型.

为检验假设 H_0, 我们要构造一个检验统计量, 使它的变化能反映 H_0 成立与否. 记

$$S_\varepsilon^2=\min_{\beta_0,\beta}\|y-\mathbf{1}\beta_0-X\beta\|^2,\ S_0^2=\min_{\beta_0}\|y-\mathbf{1}\beta_0\|^2.$$

已知

$$S_\varepsilon^2=\|P_{(1X)^\perp}y\|^2=y'P_{(1X)^\perp}y. \tag{6.34}$$

容易算出

$$S_0^2=\|P_{1^\perp}y\|^2=y'P_{1^\perp}y. \tag{6.35}$$

在 H_0 成立时计算 S_0^2 的均值有

$$\begin{aligned}E_0S_0^2&=E_0t_rP_{1^\perp}yy'=t_rP_{1^\perp}E_0yy'=t_rP_{1^\perp}[\mathrm{Cov}\,y+E_0y(E_0y)']\\&=(n-1)\sigma^2.\end{aligned}$$

在 H_0 不成立, 设原模型为真时, 计算 S_0^2 的均值得

$$E_1S_0^2=(n-1)\sigma^2+\|P_{1^\perp}X\beta\|^2>E_0S_0^2.$$

因此, 可考虑在 S_0^2 较大时拒绝 H_0. 但由于 S_0^2 的分布是 $\sigma^2\chi_{n-1}^2$, 含有未知参数, 不能作为检验统计量. 设想用 S_0^2 与 S_ε^2 之比消去 σ^2, 但二者又不独立. 为此引进回归平方和 $S_H^2=S_0^2-S_\varepsilon^2$. 易见

$$S_H^2=y'(P_{(1X)}-P_1)y\geqslant 0.\ (\text{根据附录 A.6.5})$$

因 $S_H^2+S_\varepsilon^2=S_0^2$, 由 Cochran 定理可推出

$$\sigma^{-2}S_H^2\sim\chi_p^2(\delta),\ \sigma^{-2}S_\varepsilon^2\sim\chi_{n-p-1}^2, \text{且相互独立}.$$

故得

$$F \triangleq \frac{S_H^2}{S_\varepsilon^2} \cdot \frac{n-p-1}{p} \sim F_{p, n-p-1, \delta}. \tag{6.36}$$

当 H_0 成立时

$$\delta^2 = \beta_0 \mathbf{1}'(\boldsymbol{P}_{(1X)} - \boldsymbol{P}_1)\mathbf{1}\beta_0/\sigma^2 = 0,$$

知 $F \sim F_{p, n-p-1}$.

设检验的水平为 α, $F_{p,n-p-1}(\alpha)$ 是 $F_{p,n-p-1}$ 的上侧 α–分位点，则检验 H_0 的拒绝域是

$$\{F \geqslant F_{p,n-p-1}(\alpha)\}. \tag{6.37}$$

现设假设(6.33)已被拒绝，但不见得 β 的分量全不为零. 我们还要检验因子 x_i 的效应 β_i 的显著性. 这时检验零假设

$$H_{0i}: \beta_i = 0. \tag{6.38}$$

如果经检验 H_{0i} 被拒绝，则称因子 x_i 是显著的；如果 H_{0i} 被接受，因子 x_i 就可在模型中剔除. 故称这样的检验为显著性检验.

为了以下讨论的简便，不妨假定模型是没有常数项的，即 $\beta_0 = 0$. 此时回归函数为

$$\boldsymbol{y} = \boldsymbol{X}\hat{\beta} \tag{6.39}$$

$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, $\hat{\beta}_i = \boldsymbol{e}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, 记 $\boldsymbol{c} \triangleq (\boldsymbol{X}'\boldsymbol{X})^{-1} \triangleq (c_{ij})$, 有 $\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$. 由于剩余平方和 $S_\varepsilon^2 = \boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}$ 与 $\hat{\beta}_i$ 独立，我们可令

$$F = \frac{\hat{\beta}_i^2}{c_{ii}S_\varepsilon^2}(n-p),$$

当 H_{0i} 成立，有

$$F \sim F_{1, n-p}.$$

水平为 α 的拒绝域是

$$F \geqslant F_{1, n-p}(\alpha).$$

当 H_{0i} 被接受，将 x_i 的数据剔除. 将 $\boldsymbol{X}$ 的第 i 列除去后的矩阵记为 $\boldsymbol{X}_*$, 模型就变为

$$\boldsymbol{y} = \boldsymbol{X}_*\boldsymbol{\beta}_* + \varepsilon.$$

需要重新估计 $\boldsymbol{\beta}_*$, 得新的回归函数

$$\hat{\boldsymbol{y}} = \boldsymbol{X}_*\hat{\boldsymbol{\beta}}_*. \tag{6.40}$$

不难证明，记 $\hat{\boldsymbol{\beta}}_* = (\hat{\beta}_{*1}, \cdots, \hat{\beta}_{*i-1}, \hat{\beta}_{*i+1}, \cdots, \hat{\beta}_{*p})'$, 有

下面讨论一个在挑选因子时容易引起误解的问题，并从中得出很有意思的结论.

设 $\boldsymbol{y}$, $\boldsymbol{X}$ 适合线性回归模型

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\varepsilon,\quad E\varepsilon=0 \tag{6.42}$$

如果人为地丢掉(或者客观上存在而未能用上)(6.42)中的一部分回归因子，不妨设为后面的 $p-r$ 个，是不是一定使预测的效果变坏呢？回答是不见得变坏，有时有可能更好. 让我们以在新的试验点 $\boldsymbol{x}$ 上的预测的均方误差为标准来进行讨论：

将 $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{x}$ 均进行相应的剖分为

$$\boldsymbol{X}=(\boldsymbol{X}_1^r \vdots \boldsymbol{X}_2^r),\ \boldsymbol{\beta}=\begin{pmatrix}\boldsymbol{\beta}_{(1)}\\ \boldsymbol{\beta}_{(2)}\end{pmatrix}^r,\ \boldsymbol{x}=\begin{pmatrix}\boldsymbol{x}_{(1)}\\ \boldsymbol{x}_{(2)}\end{pmatrix}^r.$$

在原模型下得到的 Y 的预测是

$$\hat{y}=\boldsymbol{x}'\hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

原模型又可改记为

$$\boldsymbol{y}=\boldsymbol{X}_1\boldsymbol{\beta}_{(1)}+\boldsymbol{X}_2\boldsymbol{\beta}_{(2)}+\varepsilon \triangleq \boldsymbol{X}_1\boldsymbol{\beta}_{(1)}+\varepsilon_*. \tag{6.43}$$

模型(6.43)中的 $E\varepsilon_*$ 实际上是 $\boldsymbol{X}_2\boldsymbol{\beta}_2$，但在丢掉这些因子时，就数据而言，我们无法去获知它的实际均值，并且在对(6.43)用最小二乘法求线性回归时，与 ε_* 的均值并无关系. 于是得 Y 在模型(6.43)下的预测是

$$\hat{y}_*=\boldsymbol{x}'_{(1)}\hat{\boldsymbol{\beta}}_{(1)},\ \hat{\boldsymbol{\beta}}_1=(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}\boldsymbol{X}'_1\boldsymbol{y}.$$

设两种情形下的预测误差分别是

$$\hat{\varepsilon}=Y-\hat{y},\ \hat{\varepsilon}_*=Y-\hat{y}_*.$$

计算均方误差得

$$\begin{aligned}E\hat{\varepsilon}^2&=D(\hat{\varepsilon})=\sigma^2(1+\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}),\\ E\hat{\varepsilon}_*^2&=D(\hat{\varepsilon}_*)+(E\hat{\varepsilon}_*)^2\\ &=\sigma^2(1+\boldsymbol{x}'_{(1)}(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}\boldsymbol{x}_{(1)})+(E\hat{\varepsilon}_*)^2,\end{aligned}$$

注意到

$$\begin{aligned}E\hat{\varepsilon}_*&=\boldsymbol{x}'\boldsymbol{\beta}-\boldsymbol{x}'_{(1)}(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}\boldsymbol{X}'_1\boldsymbol{X}\boldsymbol{\beta}\\ &=\boldsymbol{x}'_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{x}'_{(2)}\boldsymbol{\beta}_{(2)}-\boldsymbol{x}'_{(1)}(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}\boldsymbol{X}'_1(\boldsymbol{X}_1\boldsymbol{\beta}_{(1)}+\boldsymbol{X}_2\boldsymbol{\beta}_{(2)})\\ &=\boldsymbol{x}'_{(2)}\boldsymbol{\beta}_{(2)}-\boldsymbol{x}'_{(1)}(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}\boldsymbol{X}'_1\boldsymbol{X}_2\boldsymbol{\beta}_{(2)},\end{aligned}$$

得 $$(E\hat{\varepsilon}_*)^2=(\boldsymbol{x}_{(2)}-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{x}_{(1)})'\boldsymbol{\beta}_{(2)}\boldsymbol{\beta}_{(2)}'(\boldsymbol{x}_{(2)}-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{x}_{(1)}).$$

记 $E\hat{\varepsilon}_*=b$ 称作预测偏差. 由于

$$\boldsymbol{X}'\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2\\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2\end{pmatrix},\text{ 记 }(\boldsymbol{X}'\boldsymbol{X})^{-1}\triangleq\mathbf{c}\triangleq\begin{pmatrix}\mathbf{c}_{11} & \mathbf{c}_{12}\\ \mathbf{c}_{21} & \mathbf{c}_{22}\end{pmatrix}.$$

由附录 A.5.2 知, $\mathbf{c}_{11}=(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}+(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\mathbf{c}_{22}\boldsymbol{X}_2'\boldsymbol{X}_1\cdot(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$, $\mathbf{c}_{12}=-(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\mathbf{c}_{22}$. 从而得

$$\begin{aligned}D(\hat{\varepsilon})-D(\hat{\varepsilon}_*)&=\sigma^2(\boldsymbol{x}'\mathbf{c}\boldsymbol{x}-\boldsymbol{x}_{(1)}'(\boldsymbol{X}_1',\ \boldsymbol{X}_1)^{-1}\boldsymbol{x}_{(1)})\\
&=\sigma^2\boldsymbol{x}'\left(\mathbf{c}-\begin{pmatrix}(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right)\boldsymbol{x}\\
&=\sigma^2\boldsymbol{x}'\begin{pmatrix}(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\mathbf{c}_{22}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} & -\mathbf{c}_{22}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\\ -(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\mathbf{c}_{22} & \mathbf{c}_{22}\end{pmatrix}\boldsymbol{x}\\
&=\sigma^2\boldsymbol{x}'\begin{pmatrix}-(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\\ \boldsymbol{I}\end{pmatrix}\\
&\quad\times\mathbf{c}_{22}(-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{I})\boldsymbol{x}\\
&=\sigma^2(\boldsymbol{x}_{(2)}-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{x}_{(1)})'\mathbf{c}_{22}(\boldsymbol{x}_{(2)}\\
&\quad-\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{x}_{(1)}).\end{aligned}$$

故有

$$\begin{aligned}E\hat{\varepsilon}^2-E\hat{\varepsilon}_*^2=&(\boldsymbol{x}_{(2)}-(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{x}_{(1)})'(\sigma^2\mathbf{c}_{22}-\boldsymbol{\beta}_{(2)}\boldsymbol{\beta}_{(2)}')(\boldsymbol{x}_{(2)}\\
&-(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{x}_{(1)}).\end{aligned}\tag{6.44}$$

注意到 $\sigma^2\mathbf{c}_{22}$ 是正定阵, 而 $\boldsymbol{\beta}_{(2)}\boldsymbol{\beta}_{(2)}'$ 的唯一非零特征值是 $\|\boldsymbol{\beta}_{(2)}\|^2=\boldsymbol{\beta}_{(2)}'\boldsymbol{\beta}_{(2)}$. 故当 $\|\boldsymbol{\beta}_{(2)}\|$ 较小时, $\sigma^2\mathbf{c}_{22}-\boldsymbol{\beta}_{(2)}\boldsymbol{\beta}_{(2)}'$ 可以是正定阵. 此时用 $\hat{y}$ 预测 Y 不如用 $\hat{y}_*$, 因为 $E\hat{\varepsilon}^2>E\hat{\varepsilon}_*^2$.

上述讨论表明当一些因子的效应相当小时, 将它们弃置, 反而有可能改善预测的效果. 试问: 这样的改善从何而来? $\hat{y}$ 不是 Y 的均方误差最小的预测吗? 这是因为模型(6.43)已与模型(6.42)不同, $\hat{y}$ 均方误差最小是在模型为(6.42)的条件下而言的, 新模型下

的预测牺牲了无偏性，使均方误差中出现 b^2 这一项，但却使误差的方差比原来的小，二者相抵均方误差有可能比原来的小. 当然，上述讨论只是理论性的，因为在结果中含有未知参数，要得到实际可行的丢掉因子的方法，尚待进一步讨论，不在此进行。

§6.3 方差分析

(一)基本概念

方差分析是分析试验数据时最常用的统计方法。

让我们先从**因子试验模型**谈起. 设在试验(或能控制的观测)中，我们感兴趣的是结果中某个能观察到的指标 y，把它视为因变量，而影响指标 y 取值的试验条件中有若干个因子，记为 F_1，…，F_k. 例如我们感兴趣于小麦的亩产量，而影响小麦亩产的至少有土地(记为 F_1)，品种(记为 F_2)，施肥量(记为 F_3)等等因子. 在实际试验中，我们能够考察的只能是有限种不同情形，即只能在若干块土地，若干种品种和若干种施肥量的各种不同搭配下作试验. 设因子 F_i 在试验中有 S_i 种不同情形，就称 F_i 有 S_i 个水平，如施肥量分 10 斤、20 斤、30 斤三种，就是三个水平. 在因子试验模型中，我们把每个因子的每个水平各看作一个自变量(而不是把一个因子看成一个自变量，在"因子"这个词的用法上与回归分析模型有区别)，这是一个标号变量，它取值为 1 或 0，如果在某个试验点上，因子 F_i 的第 j 个水平被使用，则与之相应的自变量就取值为 1，在不被使用时就取值为零. 于是，在这个模型中，因子代表着一组有特殊联系的自变量(在各个试验点上，这组自变量必有一个取值 1，且仅仅有一个取值 1). 与自变量相应的那个参数，因为系数是 1，代表了自变量对因变量 y 的贡献，称作该自变量(也就是某因子的某个水平)的效应. 以较简单的情形为例，设土地是没有区别的，品种只有甲乙两种(设其效应分别为 β_1，β_2)，而肥料有三种水平(设其效应分别为 γ_1，γ_2，γ_3)，则就每种水平搭配各作一次试验的模型是

$$
\begin{cases}
y_{11}=\theta_0+\beta_1+\gamma_1+\varepsilon_{11}, & y_{12}=\theta_0+\beta_1+\gamma_2+\varepsilon_{12}, \\
y_{13}=\theta_0+\beta_1+\gamma_3+\varepsilon_{13}, & \\
y_{21}=\theta_0+\beta_2+\gamma_1+\varepsilon_{11}, & y_{22}=\theta_0+\beta_2+\gamma_2+\varepsilon_{22}, \\
y_{23}=\theta_0+\beta_2+\gamma_3+\varepsilon_{23}. &
\end{cases}
$$

其中 θ_0 是代表其它因素的平均效应的一个参数，用矩阵记法有

$$
\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix}=\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \theta_0 \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}+\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{pmatrix},
$$

简记为 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\theta}+\varepsilon.$ (6.45)

这里关于模型的基本假定是各水平对 y 的效应是可加的，故亦称效应可加模型. 从某种意义上来说，它比回归分析模型的条件放宽了，指的是：倘若把施肥量作为一个回归分析中的因子看待，把三个水平看作是一个自变量 x 的三个取值，那就要求 y 对 x 线性相依，否则就不能写成线性模型，但方差分析模型中把它看成三个自变量，就不存在这一要求.

在分析因子试验模型时，我们首先希望知道一个因子的各个水平效应是否相同. 如果相同，说明这个因子不管取哪种水平对指标无不同影响，那么，这个因子实际上无关紧要，可纳入平均效应中去，这时，称这个因子是不显著的；如果一个因子的各水平效应有不同，则称此因子的作用是显著的.

方差分析正是为检验因子在试验中作用的显著性而引进的一种方法. 它最早是由 R. A. Fisher 于 1920 年前后对农业试验作统计分析时引进的. 方差分析的概念如下：设在 n 个试验点上作了试验，得到观察值向量 $\boldsymbol{y}=(y_1, \cdots, y_n)'$，它满足一个线性模型 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\varepsilon$. 称 $\|\boldsymbol{y}\|^2=\boldsymbol{y}'\boldsymbol{y}=\Sigma y_\alpha^2$ 为总平方和，如果 $\|\boldsymbol{y}\|^2$ 可分解为一些非负的项的和，如 $\|\boldsymbol{y}\|^2=\Sigma\xi_i$，而这些 ξ_i 又可作出明确的统计解释，那么，在一定的条件下，上述分解式可用来进行统计推断，

则称此分解为方差分析. 一般来说就是将 $\|\boldsymbol{y}\|^2$ 分解为一些二次型的和, 如果其中某个二次型恰好反映了某个因子对 $\|\boldsymbol{y}\|^2$ 的作用, 就有可能在一定条件下用来检验该因子作用的显著性. 这里的"一定条件", 将在应用方差分析时予以说明. Cochran 定理(定理 1.2)为方差分析提供了理论基础. 在实际进行方差分析时, 常常是将 $\|\boldsymbol{P}_{1^\perp}\boldsymbol{y}\|^2=\Sigma(y_\alpha-\bar{y})^2$ 进行分解, 这里 $\bar{y}=\frac{1}{n}\sum_1^n y_\alpha$. 由于

$$\|\boldsymbol{P}_{1^\perp}\boldsymbol{y}\|^2=\|\boldsymbol{y}\|^2-n\bar{\boldsymbol{y}}^2,$$

对 $\|\boldsymbol{P}_{1^\perp}\boldsymbol{y}\|^2$ 的分解也就给出了 $\|\boldsymbol{y}\|^2$ 的分解. 由于 $\|\boldsymbol{P}_{1^\perp}\boldsymbol{y}\|^2$ 与样本方差仅差一个常数因子, 这大概是称此分解为方差分析的理由. 也有人称方差分析为离差平方和分析, 但不如方差分析简略.

方差分析方法, 自然适用于一切线性模型. 其实, 我们在上节(四)中讨论显著性检验时, 用的正是方差分析法. 但在方差分析模型中, 由于所关心的主要问题是检验因子的显著性, 寻求最佳试验条件, 方差分析的作用显得更为突出. 当然, 在方差分析模型中也需要估计, 然而此时的设计矩阵常常降秩, 在处理时还要引进一些概念和技巧, 故留待 §5 中去讨论. 对于方差分析模型来说, 试验设计问题也显示了它的重要性. 我们希望能精心安排试验, 这不仅有减少试验次数、节省支出的作用, 也是为了对数据作出明确统计解释的需要, 甚至是为了方差分析的可行性也必须对试验加以设计. 对于可以人为地控制的试验, 这也是能够做到的.

前面所谈到的类如(6.45)的模型, 被称为狭义的方差分析模型. 本节将限于讨论这类模型中的一些不算复杂的特例. 这方面的深入讨论可参看 Scheffe 的名著«The Analysis of Variance».

在以下各段中, 总假定模型是正态独立同方差的, 即假定 $\boldsymbol{\varepsilon}\sim N_n(0, \sigma^2\boldsymbol{I}_n)$.

(二)单向分类模型

单向分类模型, 又称一种方式分组模型, 也就是一个因子的试验模型. 它是方差分析模型中最简单的一种. 它就是按因子的 r 个

水平将观察值分为 r 个组，记为

$$\boldsymbol{y}=(y_{11}\cdots y_{1n_1},\ \cdots,\ y_{r1},\ \cdots,\ y_{rn_r})\triangleq(\boldsymbol{y}'_{(1)}\cdots\boldsymbol{y}'_{(r)})'.$$

其中第 i 组 $\boldsymbol{y}_{(i)}$ 是试验取第 i 个水平进行时得到的观察值向量，n_i 是在第 i 个水平上试验的重复次数，$i=1,\ \cdots,\ r$. 故称作一种方式分组数据模型，其具体结构为

$$y_{ik_i}=\beta_0+\beta_i+\varepsilon_{ik_i};\ i=1,\ \cdots,\ r;\ k_i=1,\ \cdots,\ n_i,\ \textstyle\sum_1^r n_i=n. \tag{6.46}$$

根据方差分析的思想，将观察值向量 $\boldsymbol{y}$ 的总的离差平方和 $SS_T\triangleq\sum\limits_i\sum\limits_{k_i}(y_{ik_i}-\bar{y})^2$ 进行分解. 以后，凡用 $\bar{y}$ 皆表示 $\boldsymbol{y}$ 的分量的总平均，即把所有观察值加起来，除以试验总次数. 而用 $\bar{y}_{i\cdot}$ 表示将观察值按照"·"所在位置的角标求和，然后除以这个角标的取值个数，即有 $\bar{y}_{i\cdot}=\dfrac{1}{n_i}\sum\limits_{k_i=1}^{n_i}y_{ik_i}$.

对 $\sum\sum(y_{ik_i}-\bar{y})^2$ 用平方和分解法得

$$\begin{aligned}SS_T&=\sum_i\sum_{k_i}(y_{ik_i}-\bar{y})^2=\sum_i\sum_{k_i}(y_{ik_i}-\bar{y}_{i\cdot}+\bar{y}_{i\cdot}-\bar{y})^2\\&=\sum_i\sum_{k_i}(y_{ik_i}-\bar{y}_{i\cdot})^2+\sum_i\sum_{k_i}(\bar{y}_i-\bar{y})^2.\end{aligned}\tag{6.47}$$

上式第二个等号成立是因为交叉项

$$\sum_i\sum_{k_i}(y_{ik_i}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y})=\sum_i[(\bar{y}_{i\cdot}-\bar{y})\sum_{k_i}(y_{ik_i}-\bar{y}_{i\cdot})]=0.$$

记 $SS_e=\sum\limits_i\sum\limits_{k_i}(y_{ik_i}-\bar{y}_{i\cdot})^2,\ SS_r=\sum\limits_i\sum\limits_{k_i}(\bar{y}_{i\cdot}-\bar{y})^2=\sum\limits_i n_i(\bar{y}_{i\cdot}-\bar{y})^2,$

则有

$$SS_T=SS_e+SS_r.$$

注意到 $SS_T=\|\boldsymbol{P}_{1_n^\perp}\boldsymbol{y}\|^2$，由引理 6.1 知 $SS_T/\sigma^2\sim\chi^2_{n-1,\delta}$ 相仿，因 $\sum_1^{n_i}(y_{ik_i}-\bar{y}_{i\cdot})^2=\|\boldsymbol{P}_{1_{n_i}^\perp}\boldsymbol{y}\|^2$，得 $\sigma^{-2}\|\boldsymbol{P}_{1_{n_i}^\perp}\boldsymbol{y}\|^2\sim\chi^2_{n_i-1}$，且由 $\|\boldsymbol{P}_{1_{n_i}^\perp}\boldsymbol{y}\|^2$，$i=1,\ \cdots,\ r$ 之间的相互独立性，由 Cochran 定理的系推出 $\sigma^{-2}SS_e\sim\chi^2_{n-r}$. 于是，根据(非中心情形的)Cochran 定理立得 $SS_r/\sigma^2\sim\chi^2_{r-1,\delta}$，这里 $\delta^2=\|\boldsymbol{P}_{1_n}E\boldsymbol{y}\|^2/\sigma^2$. Cochran 定理同时断言了 SS_e 与 SS_r 是独立的，从而有

$$F\triangleq\frac{SS_r}{SS_e}\cdot\frac{n-r}{r-1}\sim F_{r-1,n-r,\delta}. \tag{6.48}$$

如果 $\beta_1=\cdots=\beta_r$，$y_{i\cdot}$，$y_{j\cdot}$ 甚至 $\bar{y}$ 似应无显著差异，而当 $\beta_i\neq$

β_j，则$(\bar{y}_{i\cdot}-\bar{y})^2$，$(\bar{y}_{j\cdot}-\bar{y})^2$ 均会明显增加，因此 $SS_r=\sum_i n_i(\bar{y}_{i\cdot}-\bar{y})^2$ 是各类效应$(\beta_1, \cdots, \beta_r)$之间差异程度的衡量，当 $\beta_1=\cdots=\beta_r$ 不成立，SS_r 将偏大（严格论证可仿§6.2(四)）。且在 $\beta_1=\cdots=\beta_r$ 时易见 $\delta^2=0$。故(6.48)中 F 可用作检验假设

$$H_0:\beta_1=\cdots=\beta_r$$

的检验统计量。如果要求检验的水平为 α，则取拒绝域为

$$\{F\geqslant F_{r-1,n-r}(\alpha)\}. \tag{6.49}$$

当由观察值计算所得 F 属此拒绝域，则否定 H_0 假设，认为试验中所考察的因子是显著的。否则就认为此因子是不显著的。

实际上这个检验问题与§6.2(四)中检验回归模型的假设检验完全一致。不过这里是用典型的方差分析方法来讨论的。因设计阵是 0－1 阵，这里的计算要简单得多。

在具体进行时，把计算结果填入下面格式的方差分析表中，使方差分析成为一项很容易进行的工作。

表 6.1　单向分类方差分析表

平方和来源	平　方　和	自由度	平均平方和	F 值
类　间	$SS_r=\sum_{i=1}^{r} n_i(\bar{y}_{i\cdot}-\bar{y})^2$	$r-1$	$SS_r/r-1$	
误　差	$SS_e=\sum_{i=1}^{r}\sum_{k_i=1}^{n_i}(y_{ik_i}-\bar{y}_{i\cdot})^2$	$n-r$	$SS_e/n-r$	$F=\frac{SS_r}{SS_e}\cdot\frac{n-r}{r-1}$
总　计	$SS_T=\sum_{i=1}^{r}\sum_{k_i=1}^{n}(y_{ik_i}-\bar{y})^2$	$n-1$		

（三）两向分类模型

两向分类模型也称**两种方式分组模型**，它相当于一个两因子试验模型。(6.45)就是一个例子，既可按数据所对应的第一个因子的两个水平中的每一个而分为两类，又可按数据所对应的第二个因子的三个水平中的每一个而分为三类，一个直观的想法就是将数据设想为矩阵形式

$$\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{pmatrix},$$

那么按行分可分为两行，按列分可分为三列，行、列各算一向，就是两向分类．这样的分法当然并非任意，而是由模型的结构所决定的．一般地有

$$y_{ij}=\theta_0+\beta_i+\gamma_j+\varepsilon_{ij},\ i=1,\ \cdots,\ r;\ j=1,\ \cdots,\ c. \quad (6.50)$$

即按行分有 r 类，按列分有 c 类．称 y_{ij} 是在 $(i,\ j)$ 格中的．(6.50)是每格只做一次试验的两向分类模型．根据需要，各格中可做重复试验．如果各格中重复试验次数相等，就称这个模型是**均衡的**；否则就称为**不均衡的**．另外，由于有两个因子存在，除了各个因子单独对试验结果起作用外，还可能存在两个因子的综合作用，这种作用称为**交互作用**，这时模型中将出现**交互效应项**．这种情况下称因子水平的单独的效应为**主效应**．(6.50)就是均衡的没有交互作用的最简单的两向分类模型．

上述讨论很容易推广到多向分类模型，那时可把数据设想为多维空间的一个阵列．

下面由简到繁地讨论几种常见情形：

1．每格中只有一个试验　模型如(6.50)．要检验的假设是

$$H_{01}:\ \beta_1=\cdots=\beta_r,$$

或者是

$$H_{02}:\ \gamma_1=\cdots=\gamma_c.$$

目的就是要检验行因子或列因子（不妨如此称呼）的显著性．

仿（二）中用过的符号，且记 $\bar{y}_{\cdot j}=\dfrac{1}{r}\sum\limits_{i=1}^{r}y_{ij}$．可作如下的离差平方和分解

$$\begin{aligned} SS_T&=\sum_i\sum_j(y_{ij}-\bar{y})^2\\ &=\sum_i\sum_j(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}+\bar{y}_{i\cdot}-\bar{y}+\bar{y}_{\cdot j}-\bar{y})^2\\ &=\sum_i\sum_j(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y})^2+\sum_i c(\bar{y}_{i\cdot}-\bar{y})^2+\sum_j r(\bar{y}_{\cdot j}-\bar{y})^2\\ &\triangleq SS_e+SS_r+SS_c. \end{aligned}$$

由(二)中讨论知

$$SS_T/\sigma^2\sim\chi^2_{rc-1,\delta},\ SS_r/\sigma^2\sim\chi^2_{r-1,\delta_r},\ SS_c/\sigma^2\sim\chi^2_{c-1,\delta_c},$$

且有

$$\sigma^2\delta^2=\sum_i\sum_j(Ey_{ij}-E\bar{y})^2=\sum_i\sum_j(\alpha_i-\bar{\alpha}+\beta_j-\bar{\beta})^2$$
$$=\sum_i c(\alpha_i-\bar{\alpha})^2+\sum_j r(\beta_j-\bar{\beta})^2=\sigma^2(\delta_r^2+\delta_c^2).$$

因此有 $SS_\varepsilon/\sigma^2\sim\chi^2_{n-r-c+1}$, 并且与 SS_r, SS_c 相互独立. 从而可得

当 H_{01} 成立, $\delta_r^2=0$, $SS_r/\sigma^2\sim\chi^2_{r-1}$, 故有

$$F_1\triangleq\frac{SS_r}{SS_\varepsilon}\cdot\frac{n-r-c+1}{r-1}\sim F_{r-1,n-r-c+1}.$$

可取检验 H_{01} 的拒绝域为 $\{F_1\geqslant F_{r-1,rc-r-c+1}(\alpha)\}$. 这里 α 是检验水平.

当 H_{02} 成立, $\delta_c^2=0$, $SS_c/\sigma^2\sim\chi^2_{c-1}$, 故有

$$F_2\triangleq\frac{SS_c}{SS_\varepsilon}\cdot\frac{rc-r-c+1}{c-1}\sim F_{c-1,rc-r-c+1}.$$

可取检验 H_{02} 的拒绝域为 $\{F_2\geqslant F_{c-1,rc-r-c+1}(\alpha)\}$.

习惯上称 SS_r 是行因子引起的(离差)平方和, SS_c 是列因子引起的(离差)平方和. 称 SS_ε 为误差平方和, 在计算时它可由 $SS_T-SS_r-SS_c$ 而得. 方差分析表如下:

表 4.2　两向分类(每格一个试验)方差分析表

平方和来源	平　方　和	自由度	均　方	F　值
行因子	$SS_r=c\sum_i(\bar{y}_{i\cdot}-\bar{y})^2$	$r-1$	$SS_r/r-1$	$F_1=\frac{SS_r(c-1)}{SS_\varepsilon}$
列因子	$SS_c=r\sum_j(\bar{y}_{\cdot j}-\bar{y})^2$	$c-1$	$SS_c/c-1$	$F_2=\frac{SS_c}{SS_\varepsilon}(r-1)$
误　差	$SS_\varepsilon=SS_T-SS_r-SS_c$	$(r-1)(c-1)$	$SS_\varepsilon/(r-1)\cdot(c-1)$	
总　计	$SS_T=\sum_i\sum_j(y_{ij}-\bar{y})^2$	$rc-1$		

2. *每格中有 p 个试验情形*　现在讨论两因子试验有交互作用模型

$$y_{ijk}=\theta_0+\beta_i+\gamma_j+(\beta\gamma)_{ij}+\varepsilon_{ijk},$$
$$i=1,\cdots,r,\ j=1,\cdots,c,\ k=1,\cdots,p. \tag{6.51}$$

这里β_i, γ_j的意义同(6.50), $(\beta\gamma)_{ij}$表示行因子的第i个水平与列因子的第j个水平的交互效应，这时要检验的假设除1中已讨论过的H_{01}, H_{02}外，尚有

$$H_{03}:(\beta\gamma)_{ij}\text{对一切}(i,j)\text{均相同.}$$

如果H_{03}被接受，认为行、列因子的交互作用是不显著的(可不予考虑)；如果拒绝H_{03}，则表明交互作用显著，这时要推断哪种水平搭配最佳，就要把交互效应考虑在内.

这时的方差分析可仿(一)与(二)1. 进行. 有

$$\sum_i\sum_j\sum_k(y_{ijk}-\bar{y})^2=\sum_i\sum_j\sum_k(y_{ijk}-\bar{y}_{ij\cdot})^2+\sum_i\sum_j p(\bar{y}_{ij\cdot}-\bar{y})^2$$
$$\triangleq SS_e+SS_g,\quad \text{其中}\ \bar{y}_{ij\cdot}=\frac{1}{p}\sum_k y_{ijk}.$$

这里$SS_g=\sum\sum p(\bar{y}_{ij\cdot}-\bar{y})^2$表示各格间的离差平方和. 记$\bar{y}_{i\cdot\cdot}=\frac{1}{cp}\sum_j\sum_k y_{ijk}$, $\bar{y}_{\cdot j\cdot}=\frac{1}{rp}\sum_i\sum_k y_{ijk}$, 又有

$$SS_g=\sum_i\sum_j p(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y})^2+\sum_i pc(\bar{y}_{i\cdot\cdot}-\bar{y})^2+\sum_i pr(\bar{y}_{\cdot j\cdot}-\bar{y})^2$$
$$\triangleq SS_{rc}+SS_r+SS_c.$$

于是有总平方和

$$SS_T=SS_e+SS_{rc}+SS_r+SS_c. \tag{6.52}$$

显然有$SS_T/\sigma^2\sim\chi^2_{rcp-1,\delta}$; SS_e是由重复试验的误差引起的，称误差平方和，有$SS_e/\sigma^2\sim\chi^2_{rc(p-1)}$; $SS_r/\sigma^2\sim\chi^2_{r-1,\delta_1}$, $SS_c/\sigma^2\sim\chi^2_{c-1,\delta_2}$, 分别由行因子和列因子引起; $SS_{rc}=SS_g-SS_r-SS_c$, 它是格间平方和减去行、列因子引起的平方和，被认为是由交互作用引起的，有$SS_{rc}/\sigma^2\sim\chi^2_{(r-1)(p-1),\delta_3}$. 并且(6.52)中右边四个平方和是相互独立的.

仿1知，检验假设H_{01}, H_{02}的统计量分别是

$$F_1\triangleq\frac{SS_r}{SS_e}\cdot\frac{rc(p-1)}{r-1},\quad F_2\triangleq\frac{SS_c}{SS_e}\cdot\frac{rc(p-1)}{c-1},$$

拒绝域分别是

$$\{F_1 \geqslant F_{r-1, rc(p-1)}(\alpha)\}, \{F_2 \geqslant F_{c-1, rc(p-1)}(\alpha)\}.$$

从统计直观来讲，SS_g 表示格间的差异，如果交互作用不显著(H_{03} 成立)，它单纯由观察值的随机性及行、列因子效应所引起，应当比交互作用显著时(H_{03} 不成立)要小，故 H_{03} 不成立时，SS_g，随之又有 SS_{rc} 有偏大的趋向. (它的严格证明请看 § 6.5(四) 的引理 6.4)从而得检验 H_{03} 的统计量为

$$F_3 \triangleq \frac{SS_{rc}}{SS_\varepsilon} \cdot \frac{rc(p-1)}{(r-1)(c-1)} \sim F_{(r-1)(c-1), rc(p-1), \delta_3},$$

这里 $\delta_3^2 = \sum_i \sum_j p(E\bar{y}_{ij\cdot} - E\bar{y}_{i\cdot\cdot} - E\bar{y}_{\cdot j\cdot} + E\bar{y})^2/\sigma^2$. 当 H_{03} 成立有 $\delta_3^2 = 0$. 因此，水平为 α 的拒绝域为

$$\{F_3 \geqslant F_{(r-1)(c-1), rc(p-1)}(\alpha)\}.$$

表 4.3　两向分类(均衡情形)方差分析表

平方和来源	平方和	自由度	均方	F 值
行因子	$SS_r = cp\sum_i(\bar{y}_{i\cdot\cdot} - \bar{y})^2$	$r-1$	$SS_r/r-1$	$F_1 = \frac{SS_r}{SS_\varepsilon} \cdot \frac{rc(p-1)}{r-1}$
列因子	$SS_c = rp\sum_j(\bar{y}_{\cdot j\cdot} - \bar{y})^2$	$c-1$	$SS_c/c-1$	$F_2 = \frac{SS_c}{SS_\varepsilon} \cdot \frac{rc(p-1)}{c-1}$
交互作用	$SS_{rc} = SS_g - SS_r - SS_c$	$(r-1)(c-1)$	$SS_{rc}/(r-1)(c-1)$	$F_3 = \frac{SS_{rc}}{SS_\varepsilon} \cdot \frac{rc(p-1)}{(r-1)(c-1)}$
格间	$SS_g = p\sum_i\sum_j(\bar{y}_{ij\cdot} - \bar{y})^2$	$rc-1$		
误差	$SS_\varepsilon = SS_T - SS_g$	$rc(p-1)$	$SS_\varepsilon/rc(p-1)$	
总计	$SS_T = \sum_i\sum_j\sum_k(y_{ijk} - \bar{y})^2$	$rcp-1$		

3. 各格中试验数不全相等(非均衡情形)　考虑模型

$$y_{ijk} = \theta_0 + \beta_i + \gamma_j + (\beta\gamma)_{ij} + \varepsilon_{ijk}, \tag{6.53}$$

其中 $i=1, \cdots, r$; $j=1, \cdots, c$; 而 k 的取值是依赖于 i, j 的，有 $k = k(i, j) = 1, \cdots, n_{ij}$.

因各格中试验数不全相等，2. 中方差分析法无法直接搬用.

这时要用附加约束法(或称边界条件法). 附加约束法的可行性将在§6.5(二)的定理6.8中给出. 我们对(6.53)附加以下约束

$$\sum_i n_{\cdot i}\beta_i=0,\ \sum_j n_{\cdot j}\gamma_j=0,\ \sum_i n_{i\cdot}(\beta\gamma)_{ij}=\sum_j n_{\cdot j}(\beta\gamma)_{ij}=0, \tag{6.54}$$

这里 $n_{i\cdot}=\sum_j n_{ij},\ n_{\cdot j}=\sum_i n_{ij}$.

设要检验的假设 H_{01}, H_{02}, H_{03} 如2. 中给出, 分别用以检验行因子作用、列因子作用和交互作用的显著性. 因附加了约束(6.54), 这些假设应表述为

$$H_{01}:\ \beta_1=\cdots=\beta_r=0;$$
$$H_{02}:\ \gamma_1=\cdots=\gamma_c=0;$$
$$H_{03}:\ (\beta\gamma)_{11}=\cdots=(\beta\gamma)_{rc}=0.$$

用类似§6.2(四)中方法去导出检验统计量. 先计算 $SS_e=\min\limits_{(6.54)}\sum_i\sum_j\sum_k[y_{ijk}-\theta_0-\beta_i-\gamma_j-(\beta\gamma)_{ij}]^2$. 我们注意到在约束(6.54)下误差平方和可分解为

$$\begin{aligned}&\sum_i\sum_j\sum_k[y_{ijk}-\theta_0-\beta_i-\gamma_j-(\beta\gamma)_{ij}]^2\\&=\sum_i\sum_j\sum_k[y_{ijk}-y_{ij\cdot}+\bar{y}-\theta_0+\bar{y}_{i\cdot\cdot}-\bar{y}-\beta_i+\bar{y}_{\cdot j\cdot}-\bar{y}-\gamma_j+\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}\\&\quad-\bar{y}_{\cdot j\cdot}+\bar{y}-(\beta\gamma)_{ij}]^2=\sum_i\sum_j\sum_k[(y_{ijk}-y_{ij\cdot})^2+(\bar{y}-\theta_0)^2\\&\quad+(\bar{y}_{i\cdot\cdot}-\bar{y}-\beta_i)^2+(\bar{y}_{\cdot j\cdot}-\bar{y}-\gamma_j)^2+(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}\\&\quad+\bar{y}-(\beta\gamma)_{ij})^2].\end{aligned}\tag{6.55}$$

(在约束(6.54)下交叉项全为0.) 于是得

$$SS_e=\sum_i\sum_j\sum_k(y_{ijk}-y_{ij\cdot})^2.$$

在假设 H_{01} 及约束(6.54)下求上述误差平方和的极小值, 可通过在(6.55)中取 $\beta_i=0$ 而得

$$SS_{01}=\sum_i\sum_j\sum_k(y_{ijk}-y_{ij\cdot})^2+\sum_i\sum_j\sum_k(\bar{y}_{i\cdot\cdot}-\bar{y})^2.$$

记 $SS_r=\sum_i n_{i\cdot}(\bar{y}_{i\cdot\cdot}-\bar{y})^2$, 有检验 H_{01} 的统计量为

$$F_1 \triangleq \frac{SS_r}{SS_e} \cdot \frac{n-rc}{r-1}, \text{ 其中 } n=\sum_{i,j} n_{ij}.$$

相仿可得 $SS_c=\sum_j n_{\cdot j}(\bar{y}_{\cdot j\cdot}-\bar{y})^2$，有检验 H_{02} 的统计量为

$$F_2 \triangleq \frac{SS_c}{SS_e} \cdot \frac{n-rc}{c-1}.$$

记 $SS_{rc}=\sum_i \sum_j n_{ij}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y})^2$，有检验 H_{03} 的统计量为

$$F_3 \triangleq \frac{SS_{rc}}{SS_e} \cdot \frac{n-rc}{(r-1)(c-1)}.$$

其余讨论与 2. 完全相仿. 方差分析表也不再另外给出.

本段中所讨论的模型和方差分析法，可推广到多因子试验情形而没有任何原则性困难，但随着因子数的增加，其复杂程度将明显增加.

§6.4 协方差分析

(一)模型和统计背景

协方差分析是组合方差分析和回归分析的特征而形成的一种统计方法. 从它所处理的模型来看，其主要成分由两部分构成，如

$$\boldsymbol{y}=\boldsymbol{X}_1\boldsymbol{\beta}+\boldsymbol{X}_2\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \tag{6.56}$$

其中 $\boldsymbol{X}_1\boldsymbol{\beta}$ 部分相应的自变量是属性因子，即 $\boldsymbol{X}_1$ 是 $0-1$ 矩阵，称 $\boldsymbol{X}_1\boldsymbol{\beta}$ 为**方差分析部分**；而 $\boldsymbol{X}_2\boldsymbol{\gamma}$ 部分相应的自变量是数量因子，即 $\boldsymbol{X}_2$ 的元素允许连续取值，称 $\boldsymbol{X}_2\boldsymbol{\gamma}$ 为**回归分析部分**. 为了以后讨论方便，我们总假定

$$\mu(\boldsymbol{X}_1)\cap\mu(\boldsymbol{X}_2)=\{\boldsymbol{0}\},\ \boldsymbol{X}_2 \text{ 满列秩}. \tag{6.57}$$

一般地说(6.57)在实际问题中能够满足. 如(6.57)不满足，会给统计分析带来一些麻烦，这点可由后面的讨论看出.

就试验模型而言，(6.56)中的 $\boldsymbol{X}_1\boldsymbol{\beta}$ 往往反映试验中能够由人们精心设计和严格控制的部分，而 $\boldsymbol{X}_2\boldsymbol{\gamma}$ 则常常是试验中那部分无法由人们控制和掌握的因素的作用. 这类情况在实际问题中是极常见的.

协方差分析模型中较简单的例子，就是在单向分类模型中添加一个因子的回归项，使模型变成

$$y_{ik}=\beta_0+\beta_i+x_{ik}\gamma+\varepsilon_{ik},\quad i=1,\cdots,r,\quad k=k_i=1,\cdots,n_i. \tag{6.58}$$

这里试验次数为 $n=\sum_1^r n_i$.

R. A. Fisher 在他的名著«Statistical Methods for Research Workers»中首先提出了协方差分析的应用. 他考虑了以(6.58)为模型的实例. 设有 r 种提高茶树产量的处理，y_{ik} 表示接受了第 i 种处理的第 k 棵茶树的产量. 试验中误差的一个重要来源是某些处理碰巧被安排在产茶能力较强的一些树中. 设以 X_{ik} 记茶树在接受处理前的产量，由于茶树的相对产量说明它对年份是稳定的，X_{ik} 可用来作为茶树的产茶能力的预报因子. 添加 X_{ik} 这一因子可以调整各处理下茶树的平均产量，从而消除因茶树的产茶能力而造成的差异，使我们得到较低的试验误差，并给出较精细的各处理间的对比. 又如要比较两种饲料的配比对加速小猪增重的影响. 这时，小猪的初重应作为一个回归因子考虑在试验模型中. 这类在协方差模型中被考虑的回归中的自变量，称作**协同变量**，或称**干扰变量**. 理由是因为这类变量不是试验中能予以设计和控制的，也不是试验所关注的条件. 引进协同变量还有其它一些作用，但最通常的意义正如以上所述.

在引进了协同变量 X 之后，对模型作统计分析时将涉及 X 和 Y 的样本协方差的计算，协方差分析的名称，大致上就是由此而来的.

(二) 基本方法

在协方差分析模型(6.56)中，统计推断的主要对象是 $\boldsymbol{\beta}$. 现在我们来讨论如何运用 § 6.2 和 § 6.3 中的方法达到这一目的.

模型(6.56)可改写为

$$\boldsymbol{y}_*=\boldsymbol{X}_1\boldsymbol{\beta}+\boldsymbol{\varepsilon},$$

其中 $\boldsymbol{y}_*=\boldsymbol{y}-\boldsymbol{X}_2\boldsymbol{\gamma}$. 因 $\boldsymbol{y}_*$ 中含未知参数 $\boldsymbol{\gamma}$，不能将 $\boldsymbol{y}_*$ 看作观察值

向量．处理的办法是在 $\boldsymbol{y}_*$ 中以 $\boldsymbol{\gamma}$ 的一个适当的估计量 $\hat{\gamma}$ 代替 $\boldsymbol{\gamma}$，从而得到一个以 $\boldsymbol{z}=\boldsymbol{y}-\boldsymbol{X}_2\hat{\gamma}$ 为观察值向量的方差分析模型

$$\boldsymbol{z}=\boldsymbol{X}_1\boldsymbol{\beta}+\varepsilon_1. \tag{6.59}$$

事实上，$\hat{\gamma}$ 可从(6.56)消去方差分析部分后导出的模型

$$\boldsymbol{P}_{X_1^\perp}\boldsymbol{y}=\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2\boldsymbol{\gamma}+\boldsymbol{P}_{X_1^\perp}\boldsymbol{\varepsilon} \tag{6.60}$$

估出．(6.60)的正规方程是

$$\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2\hat{\boldsymbol{\gamma}}=\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{y}.$$

由(6.57)不难看出

$$rk\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2=rk\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2=rk\boldsymbol{X}_2-\dim(\mu(\boldsymbol{x}_1)\cap\mu(\boldsymbol{x}_2))$$
$$=rk\boldsymbol{X}_2.\quad(\text{用到附录 A.1.2 与 A.6})$$

于是有 $\hat{\boldsymbol{\gamma}}=(\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{y}$．故得

$$\boldsymbol{z}=[\boldsymbol{I}-\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}]\boldsymbol{y}.$$

现在可以看出，对模型(6.56)中方差分析部分的统计分析，可通过纯方差分析模型(6.59)进行．由于已经消去 $\boldsymbol{X}_2\boldsymbol{\gamma}$ 项，可直接应用方差分析中的现成结果，这就带来了很大方便．

一般用剩余平方和表示模型精度．对于模型(6.59)，剩余平方和为

$$\begin{aligned}SS_{\varepsilon*}&=\|\boldsymbol{P}_{X_1^\perp}\boldsymbol{z}\|^2\\&=\|[\boldsymbol{P}_{X_1^\perp}-\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}]\boldsymbol{y}\|^2\\&=\boldsymbol{y}'\boldsymbol{P}_{X_1^\perp}\boldsymbol{y}-\boldsymbol{y}'\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{y}\\&=\boldsymbol{y}'\boldsymbol{P}_{X_1^\perp}\boldsymbol{y}-\boldsymbol{y}'\boldsymbol{P}_\mu\boldsymbol{y}.\end{aligned}$$

其中 $\mu=\mu(\boldsymbol{P}_{X_1^\perp}\boldsymbol{X}_2)$．如果不引进协同变量，就相当于在模型(6.56)中令 $\boldsymbol{\gamma}=0$，这时模型的剩余平方和就是 $\boldsymbol{SS}_\varepsilon=\boldsymbol{y}'\boldsymbol{P}_{X_1^\perp}\boldsymbol{y}$．故 $\boldsymbol{y}'\boldsymbol{P}_\mu\boldsymbol{y}$ 这一项可看作引进协同变量(也就是使用协方差分析)在提高精度方面的收获．由于在对(6.59)作方差分析时必然出现如 $\boldsymbol{X}_2'\boldsymbol{P}_{X_1^\perp}\boldsymbol{y}$ 这样的量的计算，似与 $\boldsymbol{X}_2$ 和 $\boldsymbol{y}$ 的样本协方差相近，故称之为协方差分析．

在协方差分析模型(6.56)中，也可讨论回归因子的显著性检验．设 $\boldsymbol{X}_2$ 是 q 列的，可考虑检验假设

$$H_0:\gamma_{k+1}=\cdots=\gamma_q=0.$$

记 $\boldsymbol{X}_2=[\boldsymbol{X}_{21}\vdots\boldsymbol{X}_{22}]$，其中 $\boldsymbol{X}_{21}$ 为 k 列. 则假设 H_0 成立时的模型为

$$\boldsymbol{y}=\boldsymbol{X}_1\boldsymbol{\beta}+\boldsymbol{X}_{21}\boldsymbol{\gamma}_{(1)}+\boldsymbol{\varepsilon}.$$

其中 $\boldsymbol{\gamma}_{(1)}=(\gamma_1,\ \cdots,\ \gamma_k)'$. 它的剩余平方和为 $SS_0=\|\boldsymbol{P}_{(X_1\vdots X_{21})^\perp}\boldsymbol{y}\|^2$. 而模型(6.56)的剩余平方和为 $SS_e=\|\boldsymbol{P}_{(X_1\vdots X_2)^\perp}\boldsymbol{y}\|^2$. 令 $SS_H=SS_0-SS_e$，知 SS_H 与 SS_e 独立. 记 $rk\boldsymbol{X}_1=r$ 可得

$$F\triangleq\frac{SS_H}{SS_e}\cdot\frac{n-r-q}{q-k}\sim F_{q-k,n-r-q;\delta}$$

为检验统计量. 当 H_0 成立，有 $\delta=0$. 得水平为 α 的拒绝域是

$$F\geqslant F_{q-k,n-r-q}(\alpha).$$

其余问题，或者已在上两节中讨论，或者留待下节讨论，不一一赘述.

§6.5 一般线性模型的统计推断

本节将摆脱实际背景，也不涉及§6.1中的分类，从理论上抽象地讨论一般线性模型的参数统计推断. 当然，这些讨论只在本书宗旨所规定的范围内进行.

设一般线性模型为

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}. \tag{6.61}$$

其中 $\boldsymbol{X}$ 是任意的 $n\times p$ 阶实阵，$rk\boldsymbol{X}\leqslant p$，$n\geqslant p$. $\boldsymbol{\beta}$ 是未知参数向量，如不加其它约束，认为它可在整个 R^p 中取值. 对误差向量 $\boldsymbol{\varepsilon}$，至少作假定 $E\boldsymbol{\varepsilon}=\boldsymbol{0}$. 在稍后的讨论中，还需假定它的协方差结构(如 $\mathrm{Cov}\boldsymbol{\varepsilon}=\sigma^2\boldsymbol{I}$ 或 $\mathrm{Cov}\varepsilon=\sigma^2\boldsymbol{G}>0$)，并假定它的分布形式(仅讨论正态情形). 这样的线性模型是很普通的，故名称中的"一般"二字，实指"普通"而言. 我们不讨论协方差矩阵退化情形，因为那种情形下将不可避免地要用到广义逆矩阵.

(一)可估参数函数及其估计

让我们从参数向量 $\boldsymbol{\beta}$ 的最小二乘估计谈起. 按§6.2(二)中

定义，最小二乘估计是极小值问题 $\min\limits_{\beta}\|\boldsymbol{Y}-\boldsymbol{X\beta}\|^2$ 的极小值点，它是正规方程 $\boldsymbol{X'X}\hat{\boldsymbol{\beta}}=\boldsymbol{X'y}$ 的解.

为了进一步讨论最小二乘估计的性质，我们再用直观的几何方法去求解上述极小化问题. $\|\boldsymbol{y}-\boldsymbol{X\beta}\|$ 是 R^n 中 $\boldsymbol{y}$ 与 $\boldsymbol{X\beta}$ 的距离，当 $\boldsymbol{\beta}$ 在 R^p 中变化时，$\boldsymbol{X\beta}$ 张成 R^n 中线性子空间 $\mu=\mu(\boldsymbol{X})$. 极小化 $\|\boldsymbol{y}-\boldsymbol{X\beta}\|$，就是在 $\mu(\boldsymbol{X})$ 中求一向量使之与 $\boldsymbol{y}$ 的距离最小. 根据附录 A.6.2，要求的向量正是 $\boldsymbol{y}$ 到 $\mu(\boldsymbol{X})$ 的正投影，可记为 $\boldsymbol{P_X y}$. 即有

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{P_X y}. \tag{6.62}$$

此结果的几何解释见图 6.4. 不难看出 (6.62) 与正规方程是等价的. 一方面由 (6.62) 左乘 $\boldsymbol{X'}$，用正投影阵 $\boldsymbol{P_X}$ 的性质立得正规方程. 另一方面，如果 $\hat{\boldsymbol{\beta}}$ 满足正规方程，应有 $\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2=\|\boldsymbol{y}-\boldsymbol{P_X y}\|^2$. 注意到 $\|\boldsymbol{P_X y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2=\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2-\|\boldsymbol{y}-\boldsymbol{P_X y}\|^2=0$，可得 (6.62).

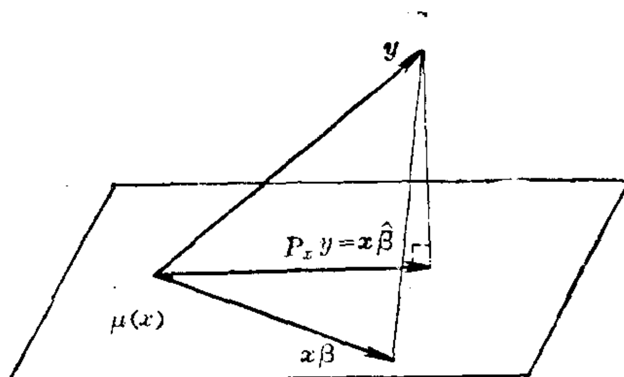


图 6.4

在 $\boldsymbol{X}$ 不满列秩 $(rk\boldsymbol{X}<p)$ 的情形下，正规方程有无穷多解，故此时谈 β 的最小二乘估计似无实际意义，习惯上就称 β 为不可估的. 为此需要引进以下概念.

定义 6.5　可估参数函数

设 $\boldsymbol{a'\beta}$ 是参数的线性函数，如果有 $\boldsymbol{y}$ 的线性函数 $\boldsymbol{c'y}$ 使得 $E\boldsymbol{c'y}=\boldsymbol{a'\beta}$，则称 $a'\beta$ 是**参数 $\boldsymbol{\beta}$ 的可估函数**，简称**可估**的；否则称 $\boldsymbol{a'\beta}$ 是**不可估**的.

我们容易证明

定理 6.4

在模型 (6.61) 中，

$$\boldsymbol{a'\beta}\ \text{可估} \Leftrightarrow \boldsymbol{a}\in\mu(\boldsymbol{X'}).$$

证　由定义 6.5 得：$\boldsymbol{a'\beta}$ 可估 $\Leftrightarrow \exists \boldsymbol{c}$ 使 $E\boldsymbol{c'y}=\boldsymbol{c'X\beta}=\boldsymbol{a'\beta}$ 对

一切 $\boldsymbol{\beta}\in R^p \Leftrightarrow \exists \mathbf{c}$ 使 $\mathbf{c}'X=\boldsymbol{a}' \Leftrightarrow \boldsymbol{a}\in\mu(X')$.

系 1° $\boldsymbol{X\beta}$ 的每个分量 $\boldsymbol{e}_i'\boldsymbol{X\beta}$, $i=1, \cdots, n$ 都是可估的，这里 $\boldsymbol{e}_i$ 是单位阵 $\boldsymbol{I}_n$ 的第 i 列，即 $\boldsymbol{e}_i=(0\cdots0\underset{i}{1}0\cdots0)'$. (在本书的后面部分，我们将使用 e_i 始终表示这一向量，而它的整个维数，可随使用的场合而定.) 2° 如 $rk\boldsymbol{X}=p$, 则对一切 $\boldsymbol{a}\in R^p$, $\boldsymbol{a}'\boldsymbol{\beta}$ 均可估；如 $rk\boldsymbol{X}<p$, 则 $\boldsymbol{\beta}$ 至少有一个分量，设为 $\beta_i=\boldsymbol{e}_i'\boldsymbol{\beta}$, 是不可估的.

证明是直接的，留作练习.

称 $rk\boldsymbol{X}<p$ 的模型为降秩的，否则为满秩的. 降秩模型中 $\boldsymbol{\beta}$ 不可估(即至少有一分量不可估)的理由可由正规方程多解性说明. 因降秩时 $\boldsymbol{X}'\boldsymbol{X\beta}=0$ 有非零解 $\boldsymbol{\beta}_0$, 则若 $\boldsymbol{\beta}_*=\boldsymbol{\beta}+\boldsymbol{\beta}_0$, 必有 $\boldsymbol{X\beta}_*=\boldsymbol{X\beta}$, 而 $\boldsymbol{\beta}$ 在模型中的作用是通过 $\boldsymbol{X\beta}$ 体现的，故从模型无从推断 $\boldsymbol{\beta}$ 和 $\boldsymbol{\beta}_*$ 的差异. 举一个简单的例子：将真实重量为 β_1 和 β_2 的物体，同时放在天平的右盘上称三次，得称重为 y_1, y_2, y_3, 于是，模型为

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}=\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}+\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}.$$

显然此模型是降秩的，$\boldsymbol{\beta}$ 不可估，其理由也可从这种称法本身不能区分 β_1 和 β_2 而得到说明.

设 $\boldsymbol{a}'\boldsymbol{\beta}$ 是可估的，记 $\boldsymbol{a}'\boldsymbol{\beta}$ 的线性无偏估计的全体为

$$\mathscr{E}_U(\boldsymbol{a})=\{\mathbf{c}'\boldsymbol{y}:\ E\mathbf{c}'\boldsymbol{y}=\boldsymbol{a}'\boldsymbol{\beta}\ \text{对一切}\ \boldsymbol{\beta}\in R^p\}.$$

定义 6.6 最优线性无偏估计

若 $\boldsymbol{b}'\boldsymbol{y}\in\mathscr{E}_U(\boldsymbol{a})$ 满足

$$D(\boldsymbol{b}'\boldsymbol{y})=\min_{\mathbf{c}'\boldsymbol{y}\in\mathscr{E}_U(\boldsymbol{a})} D(\mathbf{c}'\boldsymbol{y}) \tag{6.63}$$

则称 $\boldsymbol{b}'\boldsymbol{y}$ 是 $\boldsymbol{a}'\boldsymbol{\beta}$ 的**最优线性无偏估计**(BLUE)，也称**极小方差线性无偏估计**(MVLUE)，习惯上还称作 **Gauss–Марков 估计**(GME).

定理 6.5 (Gauss–Марков 定理)

设在模型(6.60)中假定 $\boldsymbol{\varepsilon}$ 是不相关同方差 σ^2 的(即 $\mathrm{Cov}\boldsymbol{\varepsilon}=\sigma^2\boldsymbol{I}$), $\boldsymbol{a}'\boldsymbol{\beta}$ 是可估函数，$\hat{\boldsymbol{\beta}}$ 是正规方程的任一解，则 $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{a}'\boldsymbol{\beta}$ 的

唯一的最优线性无偏估计.

证 因 $\boldsymbol{a}'\boldsymbol{\beta}$ 可估,存在 $\boldsymbol{b}\in R^n$,满足 $\boldsymbol{a}'=\boldsymbol{b}'\boldsymbol{X}$. 从而有

$$E\boldsymbol{a}'\hat{\boldsymbol{\beta}}=E\boldsymbol{b}'\boldsymbol{X}\hat{\boldsymbol{\beta}}=E\boldsymbol{b}'\boldsymbol{P}_X\boldsymbol{y}=\boldsymbol{b}'\boldsymbol{P}_X\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{b}'\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{a}'\boldsymbol{\beta}.$$

得 $\boldsymbol{a}'\hat{\boldsymbol{\beta}}\in\varepsilon_U(\boldsymbol{a})$. 任给 $\boldsymbol{c}'\boldsymbol{y}\in\varepsilon_U(\boldsymbol{a})$, 易见 $\boldsymbol{a}'=\boldsymbol{c}'\boldsymbol{X}$. 于是 $\boldsymbol{a}'\hat{\boldsymbol{\beta}}=\boldsymbol{c}'\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{c}'\boldsymbol{P}_X\boldsymbol{y}$. 这就有

$$D(\boldsymbol{c}'\boldsymbol{y})=\sigma^2\boldsymbol{c}'\boldsymbol{c}\geqslant\sigma^2\boldsymbol{c}'\boldsymbol{P}_X\boldsymbol{c}=D(\boldsymbol{c}'\boldsymbol{P}_X\boldsymbol{y})=D(\boldsymbol{a}'\hat{\boldsymbol{\beta}}).$$

得 $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ 是最优线性无偏估计. 又若上式等号成立,有 $\boldsymbol{c}'\boldsymbol{c}=\boldsymbol{c}'\boldsymbol{P}_X\boldsymbol{c}$, 可推出 $\boldsymbol{c}'(\boldsymbol{I}-\boldsymbol{P}_X)\boldsymbol{c}=\boldsymbol{c}'\boldsymbol{P}_{X^\perp}\boldsymbol{c}=0$, 从而有 $\|\boldsymbol{P}_{X^\perp}\boldsymbol{c}\|^2=0$, 得 $\boldsymbol{P}_{X^\perp}\boldsymbol{c}=0$, 即 $\boldsymbol{c}=\boldsymbol{P}_X\boldsymbol{c}$. 因此 $\boldsymbol{c}'\boldsymbol{y}=\boldsymbol{c}'\boldsymbol{P}_X\boldsymbol{y}=\boldsymbol{a}'\hat{\boldsymbol{\beta}}$. 故有 BLUE 的唯一性. 证毕.

定理 6.5 表明, 当 $\boldsymbol{a}'\boldsymbol{\beta}$ 可估($\boldsymbol{a}\in\mu(x)$), $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ 不依赖于 $\hat{\boldsymbol{\beta}}$ 是正规方程的哪个解. 因此, 不管 $\boldsymbol{X}$ 是否满列秩, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ 总是唯一的(事实上已由(6.62)保证).

附注: 如果取消 $\text{Cov}\,\boldsymbol{\varepsilon}=\sigma^2\boldsymbol{I}$ 的假定,可估性与最小二乘估计并不改变,但 $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ 就不一定是 $\boldsymbol{a}'\boldsymbol{\beta}$ 的 BLUE. 如对 $\boldsymbol{\varepsilon}$ 的协方差阵改作如下假定

$$\text{Cov}\,\boldsymbol{\varepsilon}=\sigma^2\boldsymbol{G},\ \text{这里 } \boldsymbol{G} \text{ 是已知正定阵.}$$

这时模型可记为 $\boldsymbol{y}\sim(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{G})$. 不难求出在此模型下的 BLUE. 办法是令 $\boldsymbol{z}=\boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{y}$. 原模型化为 $\boldsymbol{z}\sim(\boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{I})$. 在此模型下定理 6.5 给出 $\boldsymbol{a}'\boldsymbol{\beta}$ 的 BLUE 是 $\boldsymbol{a}'\hat{\hat{\boldsymbol{\beta}}}$,这里 $\hat{\hat{\boldsymbol{\beta}}}$ 是正规方程

$$(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})\hat{\hat{\boldsymbol{\beta}}}=\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{y} \tag{6.64}$$

的解. 称 $\hat{\hat{\boldsymbol{\beta}}}$ 是模型 $\boldsymbol{y}\sim(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{G})$的加权($\boldsymbol{G}$)的最小二乘估计. $\boldsymbol{a}'\hat{\hat{\boldsymbol{\beta}}}$ 是 $\boldsymbol{a}'\boldsymbol{\beta}$ 的 BLUE. 模型 $\boldsymbol{y}\sim(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{G})$是 Aitken 于 1935 提出的.

对于模型 $\boldsymbol{y}\sim(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{I})$. 记 $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$, 称作剩余(或残差), 并称 $\|\hat{\boldsymbol{\varepsilon}}\|^2=\|\boldsymbol{P}_{X^\perp}\boldsymbol{y}\|^2$ 为剩余平方和. 在 § 6.2 中已经提到剩余平方和除以它的自由度 $\|\hat{\boldsymbol{\varepsilon}}\|^2/n-r\,(r=rk\boldsymbol{X})$是 σ^2 的无偏估计, 记为 $\hat{\sigma}^2$.

在假定 $\varepsilon\sim N_n(0,\ \sigma^2\boldsymbol{I})$时, 估计量 $\hat{\sigma}^2$ 和 $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ 有更强的优良

性．先证明如下的

引理 6.2 对于正态模型 $\boldsymbol{y}\sim N_n(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{I})$，$T\triangleq(\boldsymbol{y}'\boldsymbol{y},\ \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X})$是关于分布族$\{P_{(\beta,\sigma^2)}\}$的充分完备统计量．这里$P_{(\beta,\sigma^2)}$是$N_n(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{I})$的分布，$\boldsymbol{\beta}\in R^p$，$\sigma^2>0$。

证 由于$P_{(\beta,\sigma^2)}$的密度函数是

$$
\begin{aligned}
f(\boldsymbol{y};\beta,\ \sigma^2)&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2\right\}\\
&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\hat{\beta}+\boldsymbol{X}\hat{\beta}-\boldsymbol{X\beta}\|^2\right\}\\
&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}[\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\beta}]\right\}\\
&\quad\cdot\exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{X\beta}\|^2\right\}.
\end{aligned}
$$

根据定理 2.3，可得$\boldsymbol{T}=(\boldsymbol{y}'\boldsymbol{y},\ \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X})$是完备充分统计量。

从而可得

定理 6.6

对于正态模型$\boldsymbol{y}\sim N_n(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{I})$，设$\boldsymbol{a}'\boldsymbol{\beta}$是可估的，则有$\boldsymbol{a}'\hat{\boldsymbol{\beta}}$和$\hat{\sigma}^2$分别是$\boldsymbol{a}'\boldsymbol{\beta}$和$\sigma^2$的唯一的一致极小方差无偏估计(UMVUE)。

证 由定理 6.5 已知$\boldsymbol{a}'\hat{\boldsymbol{\beta}}$和$\hat{\sigma}^2$分别是$\boldsymbol{a}'\boldsymbol{\beta}$和$\sigma^2$的无偏估计．

不难将$\boldsymbol{a}'\hat{\boldsymbol{\beta}}$和$\hat{\sigma}^2$表成$T$的函数．为此，先设$\boldsymbol{A}$是满足$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A}'=\boldsymbol{X}'$的矩阵(由附录 A.3.2 可验证 A 的存在性)．因$\boldsymbol{a}'\boldsymbol{\beta}$可估，有$\boldsymbol{a}'=\boldsymbol{b}'\boldsymbol{X}$，故得$\boldsymbol{a}'\hat{\boldsymbol{\beta}}=\boldsymbol{b}'\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{b}'\boldsymbol{A}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$．且有

$$\hat{\sigma}^2=\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2=\boldsymbol{y}'\boldsymbol{y}-\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{y}'\boldsymbol{y}-\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

再由引理 6.2 和定理 2.2 即得要证的结论。

(二)受约束线性模型的参数估计

在(一)中讨论可估函数$\boldsymbol{a}'\boldsymbol{\beta}$时，我们认为$\boldsymbol{\beta}$是在$R^p$中不受约束地变化的．但在实际问题中，$\boldsymbol{\beta}$不一定能无拘无束地变化，它往往受到各种因素的制约．有时为了技术上的需要(如在作非均

衡情形的方差分析时)，也对 $\boldsymbol{\beta}$ 施加人为的约束. 这类约束中，最常见的是如下的相容线性约束

$$\boldsymbol{H\beta}=\boldsymbol{\xi}$$

这里 $\boldsymbol{H}$ 是一个 $k\times p$ 阶已知矩阵. 不失一般性常设 $rk\boldsymbol{H}=k$[1].

注意到 $\boldsymbol{H}$ 有右逆 $\boldsymbol{H}_r$，即满足 $\boldsymbol{HH}_r=\boldsymbol{I}_k$. 作变换 $\boldsymbol{z}=\boldsymbol{y}-\boldsymbol{XH}_r\boldsymbol{\xi}$，$\boldsymbol{\theta}\triangleq\boldsymbol{\beta}-\boldsymbol{H}_r\boldsymbol{\xi}$，可将原模型变成 $\boldsymbol{z}\sim(\boldsymbol{X\theta},\ \sigma^2\boldsymbol{I};\ \boldsymbol{H\theta}=\boldsymbol{0})$. 这里圆括弧中最后一项表示对参数的约束. 因此，只讨论齐次约束 $\boldsymbol{H\beta}=0$ 不会有实质性的损失，却能使推导较为简洁. 下面限于讨论受约束模型

$$\boldsymbol{y}\sim(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{I};\ \boldsymbol{H\beta}=0). \tag{6.65}$$

回顾(一)中关于可估性的定理 6.4，注意到证明中用了 $\boldsymbol{\beta}$ 的任意性，故在附加约束 $\boldsymbol{H\beta}=\boldsymbol{0}$ 后，可估的充要条件变为

$$\exists\boldsymbol{c}\ \text{使得}\ \boldsymbol{c}'\boldsymbol{X\beta}=\boldsymbol{a}'\boldsymbol{\beta}\ \text{对一切}\ \boldsymbol{H\beta}=\boldsymbol{0},$$

它等价于

$$\exists\boldsymbol{c}\ \text{使得}\ \boldsymbol{c}'\boldsymbol{X\beta}=\boldsymbol{a}'\boldsymbol{\beta}\ \text{对一切}\ \boldsymbol{\beta}\in\mu^{\perp}(\boldsymbol{H}'),$$

又等价于 $\boldsymbol{X}'\boldsymbol{c}-\boldsymbol{a}\in\mu(\boldsymbol{H}')$，即

$$\boldsymbol{a}\in\mu((\boldsymbol{X}'\ \vdots\ \boldsymbol{H}'))=\mu(\boldsymbol{X}')+\mu(\boldsymbol{H}'). \tag{6.66}$$

因 $\boldsymbol{\beta}$ 满足约束 $\boldsymbol{H\beta}=\boldsymbol{0}$，$\boldsymbol{\beta}$ 的估计量 $\hat{\boldsymbol{\beta}}$ 亦理应满足 $\boldsymbol{H}\hat{\boldsymbol{\beta}}=\boldsymbol{0}$. 于是，最小二乘估计也就推广为**受约束最小二乘估计** $\hat{\boldsymbol{\beta}}_H$，它满足

$$\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_H\|^2=\min_{\boldsymbol{H}\boldsymbol{b}=0}\|\boldsymbol{y}-\boldsymbol{Xb}\|^2. \tag{6.67}$$

用平方和分解法不难求得(6.67)的解. 不妨设 $\hat{\boldsymbol{\beta}}_H$ 满足(6.67)，我们有

$$\|\boldsymbol{y}-\boldsymbol{Xb}\|^2=\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_H\|^2+\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}_H-\boldsymbol{b})\|^2 +2(\hat{\boldsymbol{\beta}}_H-\boldsymbol{b})'\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_H).$$

可知 $\hat{\boldsymbol{\beta}}_H$ 是方程

$$(\hat{\boldsymbol{\beta}}_H-\boldsymbol{b})'\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_H)=\boldsymbol{0}\quad \text{对一切}\ \boldsymbol{Hb}=\boldsymbol{0} \tag{6.68}$$

的解. 仿可估性的讨论，知(6.68)等价于

---

1) 因为凡可以被其它约束式线性表示的约束式实际上是无意义的，在删去这些无意义的约束式之后，H 就满行秩.

$$\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_H)\in\mu(\boldsymbol{H}')$$

即存在 $\boldsymbol{\lambda}$ 使得 $\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_H=\boldsymbol{H}'\boldsymbol{\lambda}$. 从而 $\hat{\boldsymbol{\beta}}_H$ 是方程

$$\begin{pmatrix}\boldsymbol{X}'\boldsymbol{X} & \boldsymbol{H}'\\ \boldsymbol{H} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{X}_H\\ \boldsymbol{\lambda}\end{pmatrix}=\begin{pmatrix}\boldsymbol{X}'\boldsymbol{y}\\ \boldsymbol{0}\end{pmatrix}\qquad(6.69)$$

的解. 称(6.69)为模型(6.65)的正规方程.

不难证明(6.69)是相容的. 事实上, 假设 p 维向量 $\boldsymbol{c}$ 和 k 维向量 $\boldsymbol{d}$ 满足

$$(\boldsymbol{c}'\quad\boldsymbol{d}')\begin{pmatrix}\boldsymbol{X}'\boldsymbol{X} & \boldsymbol{H}'\\ \boldsymbol{H} & \boldsymbol{0}\end{pmatrix}=\boldsymbol{0}$$

则有
$$(\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{d}'\boldsymbol{H}\ \vdots\ \boldsymbol{c}'\boldsymbol{H}')=\boldsymbol{0}.$$

可推出 $\boldsymbol{H}\boldsymbol{c}=\boldsymbol{0}$. 于是得

$$\boldsymbol{0}=\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c}+\boldsymbol{d}'\boldsymbol{H}\boldsymbol{c}=\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c}.$$

因而有 $\boldsymbol{X}\boldsymbol{c}=\boldsymbol{0}$. 故得

$$(\boldsymbol{c}'\quad\boldsymbol{d}')\begin{pmatrix}\boldsymbol{X}'\boldsymbol{y}\\ \boldsymbol{0}\end{pmatrix}=\boldsymbol{0}.$$

由附录 A. 3.2 知(6.69)有解.

设 $\hat{\boldsymbol{\beta}}_H$ 为(6.69)的解, 有 $\hat{\boldsymbol{\beta}}_H\triangleq\boldsymbol{L}\boldsymbol{y}$. 由 $\boldsymbol{H}\hat{\boldsymbol{\beta}}_H=\boldsymbol{0}$ 可得 $\boldsymbol{H}\boldsymbol{L}\boldsymbol{y}=\boldsymbol{0}$, 因此 $\mathrm{Cov}(\boldsymbol{H}\boldsymbol{L}\boldsymbol{y})=\sigma^2\boldsymbol{H}\boldsymbol{L}\boldsymbol{L}'\boldsymbol{H}'=\boldsymbol{0}$, 得

$$\boldsymbol{H}\boldsymbol{L}=\boldsymbol{0}.\qquad(6.70)$$

由正规方程知

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{L}\boldsymbol{y}+\boldsymbol{H}'\boldsymbol{\lambda}=\boldsymbol{X}'\boldsymbol{y}\qquad(6.71)$$

左乘 $\boldsymbol{L}'$ 推出

$$\boldsymbol{L}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{L}\boldsymbol{y}=\boldsymbol{L}'\boldsymbol{X}'\boldsymbol{y},\ \text{又有}\ \boldsymbol{L}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{L}=\boldsymbol{L}'\boldsymbol{X}',$$

得 $\boldsymbol{X}\boldsymbol{L}$ 是对称幂等阵, 即正投影阵 $\boldsymbol{P}_{XL}$. 从而有

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}_H=\boldsymbol{P}_{XL}\boldsymbol{y}.\qquad(6.72)$$

对 $\boldsymbol{X}\boldsymbol{L}$ 可给出如下引理:

引理 6.3

设 $\boldsymbol{L}\boldsymbol{y}$ 是正规方程(6.69)的任一解, $\boldsymbol{R}$ 是满足 $\mu(\boldsymbol{R})=N(\boldsymbol{H})$ 的方阵, 这里 $N(\boldsymbol{H})$ 是 $\boldsymbol{H}$ 的零空间, 则有

$$\mu(\boldsymbol{X}\boldsymbol{L})=\mu(\boldsymbol{X}\boldsymbol{R}),$$

且 $$rk(\boldsymbol{XR})=rk\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{H}\end{pmatrix}-rk\boldsymbol{H}.$$

证 因 $\boldsymbol{HR}=\boldsymbol{0}$，由(6.71)易见

$$\boldsymbol{R'X'XLy}=\boldsymbol{R'X'y},$$

可推出 $\boldsymbol{XR}=\boldsymbol{L'X'XR}=\boldsymbol{XLXR}$，得 $\mu(\boldsymbol{XR})\subset\mu(\boldsymbol{XL})$. 但是 $\mu(\boldsymbol{L})\subset N(\boldsymbol{H})=\mu(\boldsymbol{R})$，因此 $\mu(\boldsymbol{XL})\subset\mu(\boldsymbol{XR})$. 得要证之结论 $\mu(\boldsymbol{XL})=\mu(\boldsymbol{XR})$.

由附录 A.1.2 立得

$$\begin{aligned}\gamma k\boldsymbol{XR}&=\gamma k\boldsymbol{R}-\dim(\mu(\boldsymbol{R})\cap N(\boldsymbol{X}))\\&=p-\gamma k\boldsymbol{H}-\dim\left(N\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{H}\end{pmatrix}\right)\\&=p-\gamma k\boldsymbol{H}-\left(p-\gamma k\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{H}\end{pmatrix}\right)\\&=\gamma k\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{H}\end{pmatrix}-\gamma k\boldsymbol{H}\triangleq s.\end{aligned}\tag{6.73}$$

定理 6.7

设 $\boldsymbol{y}\sim(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{I},\ \boldsymbol{H\beta}=\boldsymbol{0})$，$\hat{\boldsymbol{\beta}}_H$ 是受约束最小二乘估计(不一定唯一)，$\boldsymbol{a}'\boldsymbol{\beta}$ 是可估函数，则 $\boldsymbol{a}'\hat{\boldsymbol{\beta}}_H$ 是 $\boldsymbol{a}'\boldsymbol{\beta}$ 的唯一的 BLUE.

若记 $\hat{\sigma}_H^2=\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_H\|^2/n-s$，$s$ 如(6.73)定义，则 $\hat{\sigma}_H^2$ 是 σ^2 的无偏估计.

证 因 $\boldsymbol{a}'\boldsymbol{\beta}$ 可估，必有 $\boldsymbol{d}$, $\boldsymbol{\lambda}$ 满足 $\boldsymbol{a}=\boldsymbol{X}'\boldsymbol{d}+\boldsymbol{H}'\boldsymbol{\lambda}$. 且当 $\boldsymbol{c}'\boldsymbol{y}$ 是 $\boldsymbol{a}'\boldsymbol{\beta}$ 的任一线性无偏估计，又有 $\boldsymbol{\mu}$ 使得 $\boldsymbol{a}=\boldsymbol{X}'\boldsymbol{c}+\boldsymbol{H}'\boldsymbol{\mu}$. 再注意到 $\boldsymbol{H\beta}=\boldsymbol{P}\hat{\boldsymbol{\beta}}_H=\boldsymbol{0}$. 其余证明参照定理 6.5，从而得 $\boldsymbol{a}'\hat{\boldsymbol{\beta}}_H$ 是 BLUE.

由于 $\hat{\sigma}_H^2=\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_H\|^2/n-s=\|\boldsymbol{P}_{(XL)^\perp}\boldsymbol{y}\|^2/n-s$，而 $rk\boldsymbol{XL}=s$. 易见 $\hat{\sigma}_H^2$ 是 σ^2 的无偏估计. 证毕.

下面我们讨论：如何附加约束才能使原模型不缩小？设模型为 $\boldsymbol{y}=\boldsymbol{X\beta}+\varepsilon$, $\boldsymbol{\beta}\in R^p$. 如欲加约束使模型不缩小，意味着

$$\{\boldsymbol{X\beta}:\ \boldsymbol{\beta}\in R^p\}=\{\boldsymbol{X\beta}:\ \boldsymbol{H\beta}=\boldsymbol{0}\}=\{\boldsymbol{X\beta}:\ \boldsymbol{\beta}\in N(\boldsymbol{H})\}.$$

由引理 6.3 知

$$\{X\beta;\ \beta\in N(H)\}=\{X\beta;\ \beta\in\mu(R)\}=\{XR\beta;\ \beta\in R^p\}.$$

因此,要模型不缩小,充要条件是

$$rk(XR)=rk(X'\ \vdots\ H')-rkH'=rkX',$$

它等价于

$$\mu(X')\cap\mu(H')=\{0\}. \tag{6.74}$$

设 $rkX=r<p$, 由(6.74)易见约束矩阵 H 的秩至多为 $p-r$.

在满足(6.74)的条件下附加约束, 对模型的统计推断不会发生任何实质性的影响. 从估计的角度看,我们有如下结果

定理 6.8

设模型 $y\sim(X\beta,\ \sigma^2I;\ H\beta=0)$满足条件(6.74),又设它的可估函数 $a'\beta$ 的最优线性无偏估计是 $a'\hat{\beta}_H$,则有

$$a'\hat{\beta}_H=a'\hat{\beta}, \tag{6.75}$$

这里 $\hat{\beta}$ 是模型 $y\sim(X\beta,\ \sigma^2I)$的正规方程 $X'X\hat{\beta}=X'y$ 的满足 $H\hat{\beta}=0$ 的解.

进而,当条件

$$rk((X'\ \vdots\ H'))=p \tag{6.76}$$

满足,上面的 $\hat{\beta}_H$ 和 $\hat{\beta}$ 都是唯一的.

证 由可估性条件知存在 c, μ 满足 $a'=c'X+\mu'H$. 根据(6.72)得

$$a'\hat{\beta}_H=(c'X+\mu'H)\hat{\beta}_H=c'X\hat{\beta}_H=c'P_{XL}y.$$

由条件(6.74)得 $rkXL=rk(X'\ \vdots\ H')-rkH'=rkX$. 因此 $P_{XL}=P_X$. 从而得

$$a'\hat{\beta}_H=c'P_Xy=c'X\hat{\beta}=(a'-\mu'H)\hat{\beta}=a'\hat{\beta}.$$

在条件(6.76)满足时,有

$$\begin{pmatrix} X'X & H' \\ H & 0 \end{pmatrix}\text{是满秩阵.} \tag{6.77}$$

事实上,若有

$$0=(d'\ \vdots\ f')\begin{pmatrix} X'X & H' \\ H & 0 \end{pmatrix}=(d'X'X+f'H'\ \vdots\ d'H'),$$

知 $\boldsymbol{d}'\boldsymbol{H}'=\boldsymbol{0}$, $\boldsymbol{d}'\boldsymbol{X}'=0$, 即得 $\boldsymbol{d}'(\boldsymbol{X}'\boldsymbol{X} \vdots \boldsymbol{H}')=\boldsymbol{0}$. 故由条件(6.76)立得 $\boldsymbol{d}=\boldsymbol{0}$, 从而又有 $\boldsymbol{f}=\boldsymbol{0}$, 得(6.77). 因此 $\hat{\boldsymbol{\beta}}_H$ 作为方程(6.69)的解是唯一的. 且 $\hat{\boldsymbol{\beta}}$ 作为

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} \\ \boldsymbol{H} \end{pmatrix} \hat{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}$$

的解也是唯一的.

(三)区间估计

在本段和下段中，我们限于讨论正态模型 $\boldsymbol{y} \sim N_n(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{I})$. 先给出一个关于统计量分布的结果:

定理 6.9

设 $\mathbf{c}'$ 是一个 $m \times p$ 阶矩阵，满足

$$\underset{p\times m}{\mathbf{c}} = \boldsymbol{X}' \underset{n\times m}{\boldsymbol{K}} + \boldsymbol{H}' \underset{k\times m}{\boldsymbol{S}}, \quad \boldsymbol{K}, \boldsymbol{S} \text{ 任给},$$

$\hat{\boldsymbol{\beta}}_H = \boldsymbol{Ly}$ 是(6.69)的解, $SS_{He} \triangleq \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_H\|^2$, 则有

1° $\mathbf{c}'\hat{\boldsymbol{\beta}}_H \sim N_m(\boldsymbol{K}'\boldsymbol{P}_{XL}\boldsymbol{X\beta},\ \sigma^2\boldsymbol{K}'\boldsymbol{P}_{XL}\boldsymbol{K})$;

2° $SS_{He}/\sigma^2 \sim \chi^2_{n-s}(\delta)$, 其中 $s = rk(\boldsymbol{X}' \vdots \boldsymbol{H}') - rk(\boldsymbol{H}')$, $\delta^2 = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{P}_{(XL)^\perp}\boldsymbol{X\beta}/\sigma^2$;

3° $\mathbf{c}'\hat{\boldsymbol{\beta}}_H$ 与 SS_{He} 独立.

特别，当 $\boldsymbol{H}=\boldsymbol{0}$ 时，有 $\mathbf{c}=\boldsymbol{X}'\boldsymbol{K}$, $\mu(\boldsymbol{XL})=\mu(\boldsymbol{X})$, $\hat{\boldsymbol{\beta}}_H=\hat{\boldsymbol{\beta}}$, 则有

$\mathbf{c}'\hat{\boldsymbol{\beta}} \sim N_m(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{K}'\boldsymbol{P}_X\boldsymbol{K})$, $SS_e/\sigma^2 \sim \chi^2_{n-r}$, $\mathbf{c}'\hat{\boldsymbol{\beta}}$ 与 SS_e 独立, 其中 $SS_e = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$, $r = rk\boldsymbol{X}$.

又当 $\boldsymbol{H\beta}=\boldsymbol{0}$ 满足，则有 $E\mathbf{c}'\hat{\boldsymbol{\beta}}_H = \mathbf{c}'\boldsymbol{\beta}$, $\boldsymbol{X\beta} \in \mu(\boldsymbol{XL})$, 且 $\delta^2 = 0$.

证明是直接的. 留作练习.

现在讨论区间估计问题. 对于无约束模型 $\boldsymbol{y} \sim N_n(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{I})$. 设 $\underset{m\times p\times 1}{\mathbf{c}'\boldsymbol{\beta}}$ 的各分量均可估，并且设 $rk\mathbf{c}=m$. 则有 $\boldsymbol{K}$ 满足 $\mathbf{c}=\boldsymbol{X}'\boldsymbol{K}$, 不难验证 $\boldsymbol{K}'\boldsymbol{P}_X\boldsymbol{K}$ 是满秩的. (事实上，设 $\boldsymbol{K}'\boldsymbol{P}_X\boldsymbol{K}\boldsymbol{t}=0$, 则有 $\boldsymbol{P}_X\boldsymbol{K}\boldsymbol{t}=\boldsymbol{0}$, 因此 $\boldsymbol{X}'\boldsymbol{P}_X\boldsymbol{K}\boldsymbol{t}=\boldsymbol{X}'\boldsymbol{K}\boldsymbol{t}=\mathbf{c}\boldsymbol{t}=\boldsymbol{0}$, 得 $\boldsymbol{t}=\boldsymbol{0}$.) 由定理 6.9 立即可得

$$F \triangleq \frac{(\hat{\beta}-\beta)'c(K'P_XK)^{-1}c'(\hat{\beta}-\beta)}{SS_e} \cdot \frac{n-r}{m} \sim F_{m,n-r}. \tag{6.78}$$

这是因为 $c'(\hat{\beta}-\beta)=K'X(\hat{\beta}-\beta)=K'P_X(y-X\beta)$，从而有 $(\hat{\beta}-\beta)'c(K'P_XK)^{-1}c'(\hat{\beta}-\beta)/\sigma^2 \sim \chi^2_m$，而它与 SS_e 的独立是明显的．设 $1-\alpha$ 为置信系数，则有

$$P\{F \leqslant F_{m,n-r}(\alpha)\}=1-\alpha.$$

记 R^m 中椭球

$$G(c'\hat{\beta})=\left\{z:\ (z-c'\hat{\beta})'(K'P_XK)^{-1}(z-c'\hat{\beta}) \leqslant \frac{mSS_eF_{m,n-r}(\alpha)}{n-r}\right\}$$

则有 $P\{c'\beta \in G(c'\hat{\beta})\}=1-\alpha$．因此 $G(c'\hat{\beta})$ 是 $c'\beta$ 的置信系数为 $1-\alpha$ 的置信椭球．

在 $m=1$ 的特殊情形，一般用 t-变量给出可估函数 $a'\beta$ 的区间估计．因 $a'\beta$ 可估，存在 b 满足 $a=X'b$，于是 $a'\hat{\beta} \sim N_1(a'\beta, \sigma^2 b'P_Xb)$．得

$$T \triangleq \frac{a'(\hat{\beta}-\beta)\sqrt{n-r}}{\sqrt{SS_e \cdot b'P_Xb}} \sim t_{n-r}. \tag{6.79}$$

不难算出 $a'\beta$ 的置信系数为 $1-\alpha$ 的区间估计是

$$\left[a'\hat{\beta}-\sqrt{SS_e \cdot b'P_Xb/n-r} \cdot t_{n-r}\left(\frac{\alpha}{2}\right),\right.$$
$$\left.a'\hat{\beta}+\sqrt{SS_e \cdot b'P_Xb/n-r} \cdot t_{n-r}\left(\frac{\alpha}{2}\right)\right].$$

(四) 一般线性假设检验

设模型为 $y \sim N_n(X\beta, \sigma^2 I)$．称

$$H_0:\ H\beta=0,\ (rk\underset{k\times p}{H}=k)$$

为一般线性假设．

为检验 H_0，可用似然比导出检验统计量．令似然比

$$\lambda \triangleq \frac{M}{M_H}=\frac{\max\{L(y;\beta,\sigma^2):\ \beta \in R^p,\ \sigma^2>0\}}{\max\{L(y;\beta,\sigma^2):\ H\beta=0,\ \sigma^2>0\}}, \tag{6.80}$$

其中 $$L(\boldsymbol{y};\boldsymbol{\beta},\sigma^2)=\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2\right\},$$

(已略去常数因子)记

$$SS_e=\min\{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2,\ \boldsymbol{\beta}\in R^p\},$$
$$SS_{He}=\min\{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2,\ \boldsymbol{H\beta}=0\},$$

前面已经求得

$$SS_e=\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2,\ SS_{He}=\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_H\|^2.$$

于是 $M=\max\{\sigma^{-n}\exp(-SS_e/2\sigma^2),\ \sigma^2>0\}$容易用简单的分析方法求得，它在 $\hat{\sigma}^2=\frac{SS_e}{n}$ 处达到，而

$$M=\left(\frac{SS_e}{n}\right)^{-\frac{n}{2}}\exp\left(-\frac{n}{2}\right).$$

相仿有 $$M_H=\left(\frac{SS_{He}}{n}\right)^{-\frac{n}{2}}\exp\left(-\frac{n}{2}\right).$$

从而得 $$\lambda=\frac{M}{M_H}=\left(\frac{SS_e}{SS_{He}}\right)^{-\frac{n}{2}}.$$

因 SS_e 与 SS_{He} 不独立，对 λ 予以变形得

$$\lambda=\left(\frac{SS_{He}-SS_e}{SS_e}+1\right)^{\frac{n}{2}}\triangleq(SS_H/SS_e+1)^{\frac{n}{2}}. \tag{6.81}$$

注意到 λ 是 SS_H/SS_e 的严增函数，似然比检验的否定域 $\{\lambda\geqslant c\}$ 易于用 SS_H/SS_e 表出．根据定理 6.9，我们有

$$F\triangleq\frac{SS_H}{SS_e}\cdot\frac{n-r}{r-s}\sim F_{r-s,n-r,\delta}, \tag{6.82}$$

其中 $r=\text{rk}\boldsymbol{X}$，$S=\text{rk}(\boldsymbol{X}'\,\vdots\,\boldsymbol{H}')-\text{rk}\boldsymbol{H}'$，$\delta^2=\frac{\|\boldsymbol{X\beta}-\boldsymbol{X}E\hat{\boldsymbol{\beta}}_H\|^2}{\sigma^2}$.

当假设成立，$F\sim F_{r-s,n-r}$．因此水平为 α 的拒绝域可取

$$\{F\geqslant F_{r-s,n-r}(\alpha)\}. \tag{6.83}$$

至此，一般线性假设的检验问题已在理论上得到解决．关于这一 F-检验的优良性问题，已由许宝騄教授在 1941 年解决．

下面我们讨论前几节中所用的方差分析方法在何种模型下可行的问题．

首先不难看出，上面用似然比引出的检验，也可以用方差分析来解释．我们有

$$\|\boldsymbol{y}\|^2=\|\boldsymbol{P}_{XL}\boldsymbol{y}\|^2+\|(\boldsymbol{P}_{(XL)^\perp}-\boldsymbol{P}_{X^\perp})\boldsymbol{y}\|^2+\|\boldsymbol{P}_{X^\perp}\boldsymbol{y}\|^2$$
$$\triangleq\|\boldsymbol{P}_{XL}\boldsymbol{y}\|^2+SS_H+SS_e.$$

其中 SS_e 是原模型的剩余平方和，它反映了模型的精确程度，而 SS_H 是附加约束后模型的剩余平方和与原模型的剩余平方和之差，它反映了引进约束后增加的误差情况，如果约束客观上存在，即假设 $\boldsymbol{H\beta}=\boldsymbol{0}$ 成立，SS_H 理应甚小，反之，SS_H 应偏大．在这样的统计解释下，我们可采用 F 作为检验统计量．另外，注意到 SS_H 是 χ^2 变量，当假设 H_0 成立时，它是中心的 χ^2 变量，否则为非中心 χ^2 变量．下面的引理 6.4 给出了二者的均值的比较，可作为选择单边拒绝域的理论说明．

引理 6.4

设 $\boldsymbol{y}\sim N_n(\boldsymbol{\mu},\boldsymbol{I})$，记 $\xi=\boldsymbol{y}'\boldsymbol{P}_{\mathscr{L}^\perp}\boldsymbol{y}$，$\boldsymbol{P}_{\mathscr{L}^\perp}$ 是到子空间 $\mathscr{L}$ 的正交补空间的投影阵，则 $\xi\sim\chi_d^2(\delta)$，这里 $d=n-\dim\mathscr{L}$，$\delta^2=\boldsymbol{\mu}'\boldsymbol{P}_{\mathscr{L}^\perp}\boldsymbol{\mu}$．并且有

$$E\xi=d+\boldsymbol{\mu}'\boldsymbol{P}_{\mathscr{L}^\perp}\boldsymbol{\mu}. \tag{6.84}$$

故当 $\mu\overline{\in}\mathscr{L}$ 比 $\mu\in\mathscr{L}$ 时，ξ 有偏大趋向．

证 由附录 A.6.2，知 $\boldsymbol{P}_{\mathscr{L}^\perp}=\boldsymbol{UU}'$，这里 $\boldsymbol{U}$ 满足 $\boldsymbol{U}'\boldsymbol{U}=\boldsymbol{I}_d$．记 $c_1=\boldsymbol{U}'\boldsymbol{\mu}/\delta$，作正交阵 $\boldsymbol{c}=(\boldsymbol{c}_1\cdots\boldsymbol{c}_d)$，令 $\boldsymbol{x}=\boldsymbol{c}'\boldsymbol{U}'\boldsymbol{y}$，则有 $E\boldsymbol{x}=(\delta0\cdots0)'$，得

$$\boldsymbol{x}\sim N_d((\delta0\cdots0)',\ I_d),\ \xi=\boldsymbol{x}'\boldsymbol{x}.$$

因此有 $$E\xi=Ex_1^2+\Sigma_2^dEx_i^2=d+\delta^2.$$ 证毕．

下面从方差分析的角度来讨论对设计阵 $\boldsymbol{X}$ 的要求．设模型为 $\boldsymbol{y}\sim N_n(\boldsymbol{X\beta},\sigma^2\boldsymbol{I})$．先设 $\boldsymbol{X}$ 按具体问题分为两块的情形，即 $\boldsymbol{X}=(\boldsymbol{X}_1\,\vdots\,\boldsymbol{X}_2)$，相应有 $\boldsymbol{\beta}'=(\boldsymbol{\beta}'_{(1)}\boldsymbol{\beta}'_{(2)})$．记 $\mu_i=\mu(\boldsymbol{X}_i)$，$i=1,2$．$\mu=\mu(\boldsymbol{X})$．我们给出方差分析可行的条件：

定理 6.10

$\|\boldsymbol{y}\|^2$ 可以分解为相互独立的二次型的和

$$\|\boldsymbol{y}\|^2=SS_e+SS_1+SS_2+SS_g, \tag{6.85}$$

并且SS_1和SS_2可以分别解释为由第一个因子和第二个因子引起的平方和的充要条件是

$$\mu\cap\mu_1^{\perp}\perp\mu\cap\mu_2^{\perp}. \tag{6.86}$$

证 首先有

$$\|\boldsymbol{y}\|^2=\|\boldsymbol{P}_{\mu}\boldsymbol{y}\|^2+\|\boldsymbol{P}_{\mu^{\perp}}\boldsymbol{y}\|^2, \text{记} SS_e=\|\boldsymbol{P}_{\mu^{\perp}}\boldsymbol{y}\|^2$$

设H_{0i}: $\boldsymbol{\beta}_{(i)}=0$是要检验的零假设, $i=1, 2$. 当H_{0i}成立, 模型的剩余平方和是$\|\boldsymbol{P}_{\mu_{3-i}^{\perp}}\boldsymbol{y}\|^2$, 因此

$$SS_i=\|\boldsymbol{P}_{\mu_{3-i}^{\perp}}\boldsymbol{y}\|^2-\|\boldsymbol{P}_{\mu^{\perp}}\boldsymbol{y}\|^2,\quad i=1, 2. \tag{6.87}$$

被认为是由因子i的存在对$\boldsymbol{y}$有影响而引起的, 简称为由因子i引起的平方和. 为了使因子1和因子2的影响(在模型中表现为$\boldsymbol{X}_1\boldsymbol{\beta}_{(1)}$和$\boldsymbol{X}_2\boldsymbol{\beta}_{(2)}$)与其它影响(如误差的影响, 交互作用的影响等)分离开来, 条件是SS_1, SS_2独立, 并且要使得$\|\boldsymbol{y}\|^2-SS_e-SS_1-SS_2\triangleq SS_g$是非负定二次型, 从而可应用Cochran定理, 得到假设检验所需要的统计量.

注意到$SS_1=\|(\boldsymbol{P}_{\mu}-\boldsymbol{P}_{\mu_2})\boldsymbol{y}\|^2$, 由附录A.6.5, 知$\boldsymbol{P}_{\mu}-\boldsymbol{P}_{\mu_2}=\boldsymbol{P}_{\mu\cap\mu_2^{\perp}}$, 相仿$\boldsymbol{P}_{\mu}-\boldsymbol{P}_{\mu_1}=\boldsymbol{P}_{\mu\cap\mu_1^{\perp}}$, 因此$SS_1$与$SS_2$独立的充要条件是(6.86).

由于

$$\begin{aligned}SS_g&=\|\boldsymbol{y}\|^2-\|\boldsymbol{P}_{\mu^{\perp}}\boldsymbol{y}\|^2-\|\boldsymbol{P}_{\mu\cap\mu_2^{\perp}}\boldsymbol{y}\|^2-\|\boldsymbol{P}_{\mu\cap\mu_1^{\perp}}\boldsymbol{y}\|^2\\&=\boldsymbol{y}'[\boldsymbol{P}_{\mu}-(\boldsymbol{P}_{\mu\cap\mu_2^{\perp}}+\boldsymbol{P}_{\mu\cap\mu_1^{\perp}})]\boldsymbol{y}\end{aligned}$$

由A.6.2得$\boldsymbol{P}_{\mu\cap\mu_1^{\perp}}+\boldsymbol{P}_{\mu\cap\mu_2^{\perp}}$是正投影阵(用到条件(6.86)), 它是到$\mu$的一个子空间$L$的正投影(因$\mu\cap\mu_1^{\perp}$, $\mu\cap\mu_2^{\perp}\subset\mu$), 故有$SS_g=\|\boldsymbol{P}_{\mu\ominus L}\boldsymbol{y}\|^2$. 定理得证.

系1

设$\mu(\boldsymbol{X}_1)\cap\mu(\boldsymbol{X}_2)=\{\boldsymbol{0}\}$, 则(6.86)成立的充要条件是

$$(\mu_1\triangleq)\mu(\boldsymbol{X}_1)\perp\mu(\boldsymbol{X}_2)(\triangleq\mu_2). \tag{6.88}$$

证 由(6.88)推出$\mu=\mu_1\dot{+}\mu_2$(正交直和), 于是$\mu\cap\mu_1^{\perp}=\mu_2$, $\mu\cap\mu_2^{\perp}=\mu_1$, 故有(6.86).

反之, 当(6.86)成立, 因为$\mu\cap\mu_1^{\perp}\dot{+}\mu\cap\mu_2^{\perp}\subset\mu$而$\mu=\mu\cap\mu_1^{\perp}\dot{+}\mu_1$, 故得$\mu\cap\mu_2^{\perp}\subset\mu_1$, 但是$\dim\mu\cap\mu_2^{\perp}=\dim\mu-\dim\mu_2=\dim\mu_1$

(第一个等式是因为$\mu=\mu\cap\mu_2^{\perp}\dotplus\mu_2$，后一等式由$\mu_1\cap\mu_2=\{0\}$得)，于是$\mu\cap\mu_2^{\perp}=\mu_1$. 同理有$\mu\cap\mu_1^{\perp}=\mu_2$. 得证.

系1表明，当$rk\boldsymbol{X}=p$(即$\boldsymbol{X}$满列秩时)，如果μ_1与μ_2不正交，将无法将$\|\boldsymbol{y}\|^2$分解为(6.85)中的形式，从而不能比较两个因子的显著性程度. 由于在回归分析模型中我们假定$\boldsymbol{X}$是满列秩的，因此在那种场合，为了方差分析的需要，应使$\boldsymbol{X}'\boldsymbol{X}$具有分块对角形，即

$$\boldsymbol{X}'\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{X}_2'\boldsymbol{X}_2\end{pmatrix}.$$

但对于方差分析模型，不必对设计阵有如此严格的要求，这时，只要有因子水平搭配的某种均匀性，条件(6.86)就可满足，例如对于两向分类每格中试验次数相同的模型(不妨设$\theta_0=0$)，我们有

$$\boldsymbol{X}=(\boldsymbol{X}_1\,\vdots\,\boldsymbol{X}_2)=\left(\begin{array}{ccc:c}\boldsymbol{1}_{pc} & & & \boldsymbol{1}_p \quad\quad \\ & \ddots & & \quad\ddots \quad \\ & & & \quad\quad \boldsymbol{1}_p \\ & \boldsymbol{1}_{pc} & & \vdots \\ & & \ddots & \boldsymbol{1}_p \quad\quad \\ & & & \quad\ddots \quad \\ & & \boldsymbol{1}_{pc} & \quad\quad\boldsymbol{1}_p\end{array}\right).$$
$$\underbrace{\qquad\qquad\qquad}_{r}\qquad\underbrace{\qquad\quad}_{c}$$

设$\boldsymbol{a}=(\boldsymbol{a}_{(1)}'\cdots\boldsymbol{a}_{(r)}')'\in\mu^{\perp}(\boldsymbol{X}_1)$，则有$\boldsymbol{a}_{(i)}'\boldsymbol{1}_{pc}=0$，$i=1,\cdots,r$. 而$\mu(\boldsymbol{X})$中一般元可记为

$$\boldsymbol{t}=\boldsymbol{X}\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{pmatrix}=\begin{pmatrix}\boldsymbol{1}_{pc}u_1\\ \vdots\\ \boldsymbol{1}_{pc}u_r\end{pmatrix}+\begin{pmatrix}\boldsymbol{1}_pv_1\\ \vdots\\ \boldsymbol{1}_pv_c\\ \vdots\\ \boldsymbol{1}_pv_c\end{pmatrix}$$

故当$\boldsymbol{t}\in\mu(\boldsymbol{X})\cap\mu_1^{\perp}(\boldsymbol{X})$，有

$$u_ipc+p\sum_1^c v_j=0,\ i=1,\cdots,r.$$

因此
$$\boldsymbol{t}=\mathbf{1}_n u_0+\begin{pmatrix}\mathbf{1}_p v_1\\ \vdots\\ \mathbf{1}_p v_c\\ \vdots\\ \mathbf{1}_p v_c\end{pmatrix}, cu_0=-\sum_1^c v_j,\ n=rcp.$$

相仿可求出，当 $\boldsymbol{S}\in\mu(\boldsymbol{X})\cap\mu_2^{\perp}(\boldsymbol{X})$，有

$$\boldsymbol{S}=\begin{pmatrix}\mathbf{1}_{pc}w_1\\ \vdots\\ \mathbf{1}_{pc}w_r\end{pmatrix}+\mathbf{1}_n\cdot v_0,\quad rv_0=-\sum_1^r w_i.$$

从而容易验证 $\boldsymbol{t}'\boldsymbol{s}=0$. 得条件(6.86)成立.

定理 6.10 所讨论的条件，从原则上说不难推广到 $\boldsymbol{X}=(\boldsymbol{X}_1\vdots\cdots\vdots\boldsymbol{X}_k)$ 的情形，但复杂程度将明显增加，不在此详细讨论.

因子试验中的正交设计是为了减少试验次数而提出的一种具有相当广泛的优良性的设计. 由于随着因子数的增加，如采取全面实施的试验(即对每一种水平搭配均进行试验)，试验次数会急剧增加. 为了减少试验次数，只挑选部分水平搭配做试验，称为部分实施法. 这时，为了方差分析的需要和尽可能不降低精度，采用借助于正交表设计的试验，它使得设计矩阵有某种匀称性，从而使条件(6.86)成立，同时还具有其它优良性质. 具体的设计和分析方法，可参看有关著作.

最后，在结束线性模型这一章的时候，简略地谈谈可转化为线性模型的非线性问题，从而可看出线性模型所能处理的统计问题要比凭直觉看到的更为宽广.

设 y 对 $x_1, \cdots, x_p$ 的统计依赖关系可用

$$y=F(x_1, \cdots, x_p; \beta_1, \cdots, \beta_p)+\varepsilon$$

刻划. 这里 $\beta_1, \cdots, \beta_p$ 是未知参数，ε 是随机误差. 如果存在一个可逆的连续函数 f，使得

$$f(F(x_1, \cdots, x_p; \beta_1, \cdots, \beta_p))=\sum_{i=1}^p g_i(x_1, \cdots, x_p)\varphi_i(\beta_1, \cdots, \beta_p),$$

并且满足

$$(x_1, \cdots, x_p) \mapsto (g_1, \cdots, g_p), \quad (\beta_1, \cdots, \beta_p) \mapsto (\varphi_1, \cdots, \varphi_p)$$

都是一一对应，那么，记 $z=f(y)$，$\tilde{x}_i=g_i$，$\tilde{\beta}_i=\varphi_i$，$i=1, \cdots, p$，原模型可近似地看成如下的线性模型

$$z=\sum_{i=1}^{p} \tilde{x}_i \tilde{\beta}_i+\tilde{\varepsilon}.$$

当然，在对模型作了变换之后，作严格的统计解释可能比较困难，如这里的 $\tilde{\varepsilon}$ 和原来的 ε 的关系怎样，有何种分布等，难以精确给出．但由此可近似地用线性模型的方法来处理，如给出回归直线．然后按原变换逆转，给予一定解释．这毕竟是一种实用的方法．

具体可转化的类型很多，不在此列举，仅以一例说明．如研究猪的体重(M)与其身长(L)和腰围(R)之间的关系．如从经验获知它们的关系大致为

$$M=C \cdot L^{\beta_1} \cdot R^{\beta_2}+\varepsilon,$$

可转化为 $\quad y=\log M=\log C+\beta_1 \log L+\beta_2 \log R+\tilde{\varepsilon}.$

通过观察值可得经验回归方程

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2.$$

从而以

$$\hat{M}=e^{\hat{y}}=\hat{\beta}_0 L^{\hat{\beta}_1} R^{\hat{\beta}_2}$$

作为对猪的体重的预测．

附录 A　统计中常用的矩阵代数

我们限于在实数域中讨论矩阵和线性空间．除非另有说明或在不得已的情况下，我们总用黑体字表示矩阵或向量，而用英文花体字母表示线性空间，但 n 维欧氏空间按习惯仍用 R^n 表示．

A.1

设 $\boldsymbol{A}=(a_{ij})$ 是 $m \times n$ 阶矩阵．需要注明阶数时记为 $\underset{m\times n}{\boldsymbol{A}}$ 或 $(a_{ij})_{m\times n}$．记 $\boldsymbol{A}$ 的第 j 列向量为

$$\boldsymbol{a}_j=\begin{pmatrix}a_{1j}\\ \vdots\\ a_{mj}\end{pmatrix},\quad 有\ \boldsymbol{a}_j\in R^m,\ j=1,\ \cdots,\ n.$$

由 $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n$ 张成的 R^m 中的线性(子)空间

$$\mu(\boldsymbol{A})=\{\sum_1^n x_j\boldsymbol{a}_j\colon x_j\in R',\ j=1,\ \cdots,\ n\}$$

称作 $\boldsymbol{A}$ 的列空间.

将 $\boldsymbol{A}$ 看作 R^n 到 R^m 的线性变换，列空间 $\mu(\boldsymbol{A})$ 就是 $\boldsymbol{A}$ 的值域 $\{\boldsymbol{A}\boldsymbol{x}\colon \boldsymbol{x}\in R^n\}$. 令

$$N(\boldsymbol{A})=\{\boldsymbol{x}\colon \boldsymbol{x}\in R^n,\ \boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}\}.$$

称 $N(\boldsymbol{A})$ 为 $\boldsymbol{A}$ 的零空间(或 $\boldsymbol{A}$ 的核).

$\boldsymbol{A}$ 的秩就是 $\mu(\boldsymbol{A})$ 的维数，即

$$rk\boldsymbol{A}=\dim\mu(\boldsymbol{A}).$$

线性空间理论中一个基本事实是

A.1.1 $n=\dim\mu(\boldsymbol{A})+\dim N(\boldsymbol{A})$.

注意到只要 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 可乘，就有

$$\mu(\boldsymbol{A}\boldsymbol{B})\subset\mu(\boldsymbol{A}).$$

由 A.1.1 不难推出

A.1.2 $rk\boldsymbol{A}\boldsymbol{B}=rk\boldsymbol{B}-\dim(\mu(\boldsymbol{B})\cap N(\boldsymbol{A}))$.

当 $rk\underset{m\times n}{\boldsymbol{A}}=n$，称 $\boldsymbol{A}$ 满列秩，当 $rk\underset{m\times n}{\boldsymbol{A}}=m$，称 $\boldsymbol{A}$ 满行秩. $\boldsymbol{I}_p$ 表示 p 阶单位阵. 我们有

A.1.3 (i) $\boldsymbol{A}$ 满列秩 $\Leftrightarrow$ 存在 $\boldsymbol{A}_l$ 满足 $\boldsymbol{A}_l\boldsymbol{A}=\boldsymbol{I}_n$，称 $\boldsymbol{A}_l$ 为 $\boldsymbol{A}$ 的左逆; (ii) $\boldsymbol{A}$ 满行秩 $\Leftrightarrow$ 存在 $\boldsymbol{A}_r$ 满足 $\boldsymbol{A}\boldsymbol{A}_r=\boldsymbol{I}_m$，称 $\boldsymbol{A}_r$ 为 $\boldsymbol{A}$ 的右逆.

A.2

设 $\mathscr{L}$, $\mathscr{L}_1$, $\mathscr{L}_2$ 是 R^n 的子空间. 定义 $\mathscr{L}_1$ 与 $\mathscr{L}_2$ 的和空间和列空间分别为

$$\mathscr{L}_1+\mathscr{L}_2=\{\alpha\boldsymbol{u}+\beta\boldsymbol{v}\colon \alpha,\ \beta\in R^1,\ \boldsymbol{u}\in\mathscr{L}_1,\ \boldsymbol{v}\in\mathscr{L}_2\},$$

$$\mathscr{L}_1\cap\mathscr{L}_2=\{\boldsymbol{u}\colon \boldsymbol{u}\in\mathscr{L}_1\ 且\ \boldsymbol{u}\in\mathscr{L}_2\}.$$

当 $\mathscr{L}_1\cap\mathscr{L}_2=\{\boldsymbol{0}\}$ ($\boldsymbol{0}$ 为零向量)，称 $\mathscr{L}_1+\mathscr{L}_2$ 为直和，记为 $\mathscr{L}_1\oplus$

$\mathscr{L}_2$. 因为有

A.2.1 $\dim(\mathscr{L}_1+\mathscr{L}_2)=\dim\mathscr{L}_1+\dim\mathscr{L}_2-\dim(\mathscr{L}_1\cap\mathscr{L}_2)$. 推论得

$$\dim(\mathscr{L}_1\oplus\mathscr{L}_2)=\dim\mathscr{L}_1+\dim\mathscr{L}_2.$$

如果 $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \triangleq \boldsymbol{x}_2'\boldsymbol{x}_1=0 \quad \forall \boldsymbol{x}_i\in\mathscr{L}_i,\ i=1, 2,$
则称 $\mathscr{L}_1$ 与 $\mathscr{L}_2$ 正交,记为 $\mathscr{L}_1\perp\mathscr{L}_2$.

A.2.2 $\mathscr{L}_1\perp\mathscr{L}_2\Rightarrow\mathscr{L}_1+\mathscr{L}_2=\mathscr{L}_1\oplus\mathscr{L}_2$ 称为正交直和，或记为 $\mathscr{L}_1\dot{+}\mathscr{L}_2$.

$\mathscr{L}$ 的正交补(空间)是

$$\mathscr{L}^{\perp}=\{\boldsymbol{u}: \boldsymbol{u}\in R^n,\ \boldsymbol{u}\perp\mathscr{L}\}.$$

A.2.3 $\mathscr{L}\dot{+}\mathscr{L}^{\perp}=R^n$. 即 $\mathscr{L}$ 与 $\mathscr{L}^{\perp}$ 互为正交补空间.

A.2.4 $\mathscr{L}_1\subset\mathscr{L}_2\Leftrightarrow\mathscr{L}_1^{\perp}\supset\mathscr{L}_2^{\perp}$.

A.3

线性方程 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 相容(有解)的充要条件是下列之一:

A.3.1 $\boldsymbol{b}\in\mu(\boldsymbol{A})$;

A.3.2 $\boldsymbol{c}'\boldsymbol{A}=\boldsymbol{0}\Rightarrow\boldsymbol{c}'\boldsymbol{b}=\boldsymbol{0}$;

A.3.3 $rk(\boldsymbol{A}\,\vdots\,\boldsymbol{b})=rk\boldsymbol{A}$.

A.4

设 $\boldsymbol{A}$ 是 n 阶对称方阵,则 $\boldsymbol{A}$ 的特征值都是实数，将它们从大到小排列为

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n.$$

A.4.1 设 $\boldsymbol{A}'=\boldsymbol{A}$,存在正交阵 $\boldsymbol{C}$ 满足

$$\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C}=\boldsymbol{\Lambda}\triangleq\mathrm{diag}(\lambda_1, \cdots, \lambda_n).$$

这里 $\mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ 是以 $\lambda_1, \cdots, \lambda_n$ 为顺序主对角元的对角阵. 并且,$\boldsymbol{c}_j$($\boldsymbol{C}$ 的第 j 列)是 $\boldsymbol{A}$ 的相应于 λ_j 的(单位)特征向量, $j=1, \cdots, n$. 此时易见 $\boldsymbol{A}$ 有谱分解

$$\boldsymbol{A}=\textstyle\sum_1^n\lambda_j\boldsymbol{c}_j\boldsymbol{c}_j'.$$

A.4.2 设 $\boldsymbol{A}'=\boldsymbol{A}$, $\boldsymbol{A}$ 的特征值有如下极值表示

$$\lambda_1=\max_{\|x\|=1} x'Ax, \text{在 } x=c_1 \text{ 达到.}$$

$$\lambda_i=\max_{\|x\|=1,x\perp\mu(c_1,\cdots,c_{i-1})} x'Ax, \text{在 } x=c_i \text{ 达到}, i=2,\cdots,n.$$

A.4.3 设 $A'=A$, $\underset{n\times p}{U}$ 是列正交阵, 即 U 满足 $U'U=I_p$. 则有

$$\max_{U'U=I_p} t_r U'AU=\sum_1^p\lambda_i, \text{在 } U=(c_1\cdots c_p) \text{ 达到.}$$

以上 c_i 均指 A 的第 i 个特征向量.

A.4.4 设 $A'=A$, 下列命题等价

(i) A 的特征值 $\lambda_1\geqslant\cdots\geqslant\lambda_n\geqslant 0$;

(ii) $x'Ax\geqslant 0\ \forall x\in R^n,\ x\neq 0$;

(iii) 记 $r=rkA$, 存在矩阵 B 使得 $A=BB'$. 此时, 称 A 为非负定阵, 记作 $A\geqslant 0$.

在上述(i)、(ii)中将 $\geqslant 0$ 改为 >0, 在(iii)中附加 $rkB=n$, 则得正定阵 A 的刻划, 记为 $A>0$.

A.4.5 $A\geqslant 0\Rightarrow$ 存在唯一的非负定阵

$$A^{\frac{1}{2}}\triangleq C\Lambda^{\frac{1}{2}}C'$$

使得 $A=A^{\frac{1}{2}}\cdot A^{\frac{1}{2}}$. 这里 C 如 A.4.1.

A.4.6 $\underset{n\times m}{BB'}=\underset{n\times p}{DD'}(p\leqslant m)\Leftrightarrow B=(C\,\vdots\,0)U$, 其中 U 是 m 阶正交阵.

A.4.7 $rkBB'=rkB$.

A.5

将 $\underset{m\times n}{A}$ 和 $\underset{n\times k}{B}$ 剖分为

$$A=\begin{pmatrix}A_{11}&\cdots&A_{1q}\\ \vdots&&\vdots\\ A_{p1}&\cdots&A_{pq}\end{pmatrix},\quad B=\begin{pmatrix}B_{11}&\cdots&B_{1r}\\ \vdots&&\vdots\\ B_{q1}&\cdots&B_{qr}\end{pmatrix}$$

并且满足 A_{il} 和 B_{lj} 可乘, $i=1,\cdots,p$; $j=1,\cdots,r$; $l=1,\cdots,q$. 则有

$$\boldsymbol{AB} \triangleq \boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \cdots & \boldsymbol{C}_{1r} \\ \vdots & & \vdots \\ \boldsymbol{C}_{p1} & \cdots & \boldsymbol{C}_{pr} \end{pmatrix}$$

其中 $\quad \boldsymbol{C}_{ij} = \sum_{l=1}^{q} \boldsymbol{A}_{il}\boldsymbol{B}_{lj}, \ i=1, \cdots, p; j=1, \cdots, r.$

A.5.1 设

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}, \ \boldsymbol{A}_{ii} \text{ 是 } p_i \text{ 阶方阵}, i=1, 2.$$

(i) 如 $\boldsymbol{A}_{11}$ 可逆, 有 $\det\boldsymbol{A} = \det\boldsymbol{A}_{11} \cdot \det(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})$;

(ii) 如 $\boldsymbol{A}_{22}$ 可逆, 有 $\det\boldsymbol{A} = \det\boldsymbol{A}_{22} \cdot \det(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})$.

当 $\boldsymbol{A}_{11}$ 可逆, 有

$$\begin{pmatrix} \boldsymbol{I}_{p1} & \boldsymbol{0} \\ -\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{I}_{p2} \end{pmatrix} \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{p1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \\ \boldsymbol{0} & \boldsymbol{I}_{p2} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \end{pmatrix}$$

故当 $\boldsymbol{A}$ 可逆时有

$$\boldsymbol{A}^{-1} = \left[\begin{pmatrix} \boldsymbol{I}_{p1} & \boldsymbol{0} \\ \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{I}_{p2} \end{pmatrix} \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{I}_{p1} & \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \\ \boldsymbol{0} & \boldsymbol{I}_{p2} \end{pmatrix} \right]^{-1}$$

A.5.2

记 $\quad \boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{pmatrix}$, 有

$$\boldsymbol{A}^{11} = \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}$$

$$\boldsymbol{A}^{12} = -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1}$$

$$\boldsymbol{A}^{21} = -(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}$$

$$\boldsymbol{A}^{22} = (\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1}$$

当 $\boldsymbol{A}_{22}$ 可逆时, 有类似结论.

A.5.3 设 $\boldsymbol{A} > \boldsymbol{0}$, 剖分如 A.5.1, $\boldsymbol{A}^{-1}$ 如 A.5.2, 则有 $\boldsymbol{A}^{11} - \boldsymbol{A}_{11}^{-1} \geqslant \boldsymbol{0}$, 且有

$$A^{-1}-\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}=\begin{pmatrix} -A_{11}^{-1}A_{12} \\ I \end{pmatrix} A^{22}\begin{pmatrix} -A_{11}^{-1}A_{12} \\ I \end{pmatrix}' \geqslant 0.$$

A.5.4 设 $\boldsymbol{A}$ 是 $m\times n$ 阶阵，$\boldsymbol{B}$ 是 $n\times m$ 阶阵. 则由 A.5.1 计算

$$\det\begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{B} \\ \boldsymbol{A} & \lambda\boldsymbol{I}_m \end{pmatrix} \quad (\lambda\neq 0)$$

可得 $\det(\lambda\boldsymbol{I}_m-\boldsymbol{AB})=\lambda^{-n}\det(\lambda\boldsymbol{I}_n-\boldsymbol{BA})$. 因此

$\boldsymbol{AB}$ 与 $\boldsymbol{BA}$ 有相同的(包括重数)非零特征值.

A.6

设 $\mathscr{L}\subset R^n$ 是线性空间. 如果 n 阶方阵 $\boldsymbol{P}_{\mathscr{L}}$ 满足

$$\boldsymbol{P}_{\mathscr{L}}\boldsymbol{x}=\boldsymbol{x} \quad \forall \boldsymbol{x}\in\mathscr{L},\ \boldsymbol{P}_{\mathscr{L}}\boldsymbol{y}=\boldsymbol{0} \quad \forall \boldsymbol{y}\in\mathscr{L}^{\perp},$$

则称 $\boldsymbol{P}_{\mathscr{L}}$ 是到 $\mathscr{L}$ 的正投影阵.

A.6.1 设 $\boldsymbol{A}$ 是满列秩阵，使得 $\mu(\boldsymbol{A})=\mathscr{L}$, 则有到 $\mathscr{L}$ 的正投影阵

$$P_{\mathscr{L}}=\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'.$$

设 $\boldsymbol{B}$ 满足 $\mu(\boldsymbol{B})=\mathscr{L}$, 可记 $\boldsymbol{P}_{\mathscr{L}}\triangleq\boldsymbol{P}_B$.

A.6.2 $\boldsymbol{P}$ 是到 $\mu(\boldsymbol{P})$ 的正投影阵的充要条件是下列命题中的任一个：

(i) $\boldsymbol{P}$ 是对称幂等阵;

(ii) $\|\boldsymbol{x}-\boldsymbol{Px}\|^2=\min\limits_{t\in R^n}\|\boldsymbol{x}-\boldsymbol{Pt}\|^2 \quad \forall \boldsymbol{x}\in R^n$ 成立;

(iii) $\boldsymbol{P}^2=\boldsymbol{P}$, 并且有 $\|\boldsymbol{Px}\|\leqslant\|\boldsymbol{x}\| \quad \forall \boldsymbol{x}\in R^n$;

(iv) 存在正交阵 $\boldsymbol{U}$ 使得

$$\boldsymbol{P}=\boldsymbol{U}\begin{pmatrix} \boldsymbol{I}_r & 0 \\ 0 & 0 \end{pmatrix}\boldsymbol{U}',\ r=rk\boldsymbol{P}.$$

记 $\boldsymbol{U}=(\boldsymbol{U}_1^{r} \vdots \boldsymbol{U}_2)$, 得 $\boldsymbol{P}=\boldsymbol{U}_1\boldsymbol{U}_1'$, 且有 $\boldsymbol{U}_1'\boldsymbol{U}_1=\boldsymbol{I}_r$. (即 $\boldsymbol{U}_1$ 的 r 个列向量构成了 $\mu(\boldsymbol{P})$ 的正交规范基.)

(v) $\boldsymbol{I}-\boldsymbol{P}$ 是对称幂等阵.

A.6.3 设 $\boldsymbol{P}_A$ 是到 $\mu(\boldsymbol{A})$ 的投影阵, 则 $\boldsymbol{I}-\boldsymbol{P}_A$ 是到 $\mu^{\perp}(\boldsymbol{A})$ 的投影阵, 常简为 $\boldsymbol{P}_{A^{\perp}}$.

A.6.4 设 $\boldsymbol{A}$, $\boldsymbol{B}$ 是 n 阶对称阵, $\boldsymbol{C}=\boldsymbol{A}+\boldsymbol{B}$ 幂等, 则有

(i) $\boldsymbol{A}$, $\boldsymbol{B}$ 是幂等阵 $\Leftrightarrow \boldsymbol{AB}=\boldsymbol{0}$,

(ii) $\boldsymbol{A}$ 幂等, $\boldsymbol{B}\geqslant\boldsymbol{0}\Rightarrow\boldsymbol{B}$ 幂等.

A.6.5 设 $\mathscr{L}_1$, $\mathscr{L}_2$ 是 R^n 的线性子空间, 则下列命题等价:

(i) $\mathscr{L}_1\supset\mathscr{L}_2$;

(ii) $\boldsymbol{P}_{\mathscr{L}_1}-\boldsymbol{P}_{\mathscr{L}_2}\geqslant 0$;

(iii) $\boldsymbol{P}_{\mathscr{L}_1}\boldsymbol{P}_{\mathscr{L}_2}=\boldsymbol{P}_{\mathscr{L}_2}$;

(iv) $\boldsymbol{P}_{\mathscr{L}_2}\boldsymbol{P}_{\mathscr{L}_1}=\boldsymbol{P}_{\mathscr{L}_2}$.

且当上述任一条件满足时,

$$\boldsymbol{P}_{\mathscr{L}_1}-\boldsymbol{P}_{\mathscr{L}_2}=\boldsymbol{P}_{\mathscr{L}_1\cap\mathscr{L}_2^{\perp}}$$

是到 $\mathscr{L}_1\cap\mathscr{L}_2^{\perp}$ 的正投影阵.

对于本附录的证明, 可参看倪国熙《常用的矩阵理论和方法》.

习　　题

1. 假设我们观察模型

$$y_i=\beta_0+\beta_1x_i+\beta_2x_i^2+\varepsilon_i.$$

其中 ε_i 独立且遵从 $N(0,\ \sigma^2)$. 得观察数据

x	1	2	3	4	5
y	2	8	9	5	6

试求 $\beta=(\beta_0,\ \beta_1,\ \beta_2)'$ 与 σ^2 的极大似然估计.

2. 设 $\boldsymbol{y}=(y_1,\ y_2,\ y_3)'$ 满足线性模型 $(\mathbf{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{I}_3)$,

$$E(y_i)=\beta_0+\beta_1x_i+\beta_2(3x_i^2-2),\ i=1,\ 2,\ 3.$$

其中 $x_1=-1$, $x_2=0$, $x_3=1$. 写出 X 并求 $(\beta_0,\ \beta_1,\ \beta_2)'$ 的最小二乘估计.

3. 某职工医院用光电比色计检验尿汞时, 得尿汞含量(mg/l)与消光系数读数的结果如下:

尿汞含量 (x_i)	2	4	6	8	10
消光系数 (y_i)	64	138	205	285	360

已知它们之间服从线性模型

$$Ey_i=\beta_0+\beta_1x_i$$

试求 β_0 和 β_1 的最小二乘估计, 并检验 β_1 是否为零 $(\alpha=0.05)$?

4．下面记录了一批婴儿出生体重与6个月时体重的数据，试估计它们之间的相关系数和回归直线．

人　数	6个月时体重(kg)									合计
	5.5—	6.0—	6.5—	7.0—	7.5—	8.0—	8.5—	9.0—	9.5—	
出　2.50—		1	1	1	2					5
身　2.75—		2	4	5	2	2	1			16
体　3.00—		2	2	4	8	4	2			22
重　3.25—	1		2	6	2	2	2		1	16
(kg)3.50—			1		4	2	3	1	1	12
3.75—				1		4	1	1	1	8
4.00—							1	1		2
合　计	1	5	10	17	18	14	10	3	3	

5．对三种不同品种的小麦，施用四种不同的肥料所获得的产量(每块按磅计)如下表．

	不　同　的　小　麦		
肥　料	A	B	C
α	8	3	7
β	10	4	8
γ	6	5	6
δ	8	4	7

试检验三种不同小麦的平均产量相等的假设和四种肥料等效应的零假设．

6．下面记录了三位操作工分别在四台不同机器上操作三天的日产量：

机　器	操　作　工		
	甲	乙	丙
M_1	15, 15, 17	19, 19, 16	16, 18, 21
M_2	17, 17, 17	15, 15, 15	19, 22, 22
M_3	15, 17, 16	18, 17, 16	18, 18, 18
M_4	18, 20, 22	15, 16, 17	17, 17, 17

试检验：(1) 操作工之间的差异是否显著？

(2) 机器之间的差别是否显著？

(3) 交互影响是否显著($\alpha=0.05$)？

7. 设 $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$

求出 Y 对 X 的回归 $y=E(Y/x)$ 和 X 对 Y 的回归 $x=E(X|Y)$. 观察这两条回归直线是否一致.

8. 线性回归模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, $\mathrm{Var}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}$, $\boldsymbol{X}$ 为 $n\times p$ 矩阵，秩为 p. 证明：在 m 个试验点的预测误差的方差之和为 $(m+p)\sigma^2$.

9. 试证引理 6.1 的必要性. (提示：证(i)的必要性时要用特征函数. 证(ii)的必要性时要注意到利用 B 的线性无关的列.)

10. 设总体满足模型 $y=\beta_0+\beta'\boldsymbol{X}+\varepsilon$, 在试验点 $(x_{\alpha 1}, \cdots, x_{\alpha p})$ 上得到观察值 y_α, $\alpha=1, \cdots, n$. 证明：在(6.18)中以样本矩代 $\beta_0 \triangleq EY-\boldsymbol{\sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}E\boldsymbol{X}$, $\beta \triangleq \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XY}$ 中的总体矩，所得结果记为 $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}$, 与(6.28)中用最小二乘法所得回归系数 $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}$. 完全一致.

11. 试证：在回归分析中对于假设(6.38)的检验，与自变量因子的尺度无关，即若设计矩阵的某列乘以常数 a, 检验统计量 $F=\hat{\boldsymbol{\beta}}_i^2/(c_{ii}s_i^2)\cdot(n-p)$ 并不改变.

12. 证明定理 6.4 的系.

13. 证明定理 6.9.

14. 试证：在模型(6.51)中欲使参数函数 $\sum_{i=1}^{r} b_i\beta_i$ 可估，充要条件是 $\sum_{i=1}^{r} b_i=0$. (称这样的参数函数为因子水平效应的对照.)

15. 证明(6.64)给出的 $\hat{\hat{\boldsymbol{\beta}}}$ 是模型 $\boldsymbol{y}\sim(\boldsymbol{X\beta}, \sigma^2\boldsymbol{G})$ 中 $\boldsymbol{\beta}$ 的 GM 估计. 因此，一般地说，在模型 $\boldsymbol{y}\sim(\boldsymbol{X\beta}, \sigma^2\boldsymbol{G})$ 中 $\boldsymbol{\beta}$ 的最小二乘估计 $\hat{\boldsymbol{\beta}}$ 不再是 BLUE. 如果假定 $\boldsymbol{X}$ 满列秩，你能给出此时的 $\hat{\boldsymbol{\beta}}$ 仍是 BLUE 的充要条件吗(用 $\boldsymbol{X}$, $\boldsymbol{G}$ 表出)？

16. 设 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, $\mathrm{Cov}\boldsymbol{\varepsilon}=\sigma^2\boldsymbol{G}$, 证明：

(1) $E\boldsymbol{\varepsilon}'\boldsymbol{A}\boldsymbol{\varepsilon}=\sigma^2\mathrm{tr}\boldsymbol{AG}$

(2) 当 $\boldsymbol{\varepsilon}$ 正态，有 $D(\boldsymbol{\varepsilon}'\boldsymbol{A}\boldsymbol{\varepsilon})=2\sigma^4\mathrm{tr}\boldsymbol{AGAG}$.

17. 把下面两个估计问题纳入线性模型的范围：(a) $X_1, \cdots, X_n$ 是从 $N(a, \sigma^2)$ 中的 iid. 样本，估计 a. (b) $X_1, \cdots, X_m$ 和 $Y_1, \cdots, Y_n$ 是从 $N(a, \sigma^2)$ 和 $N(b, \rho\sigma^2)$ 中取出的 iid. 样本，ρ 已知，估计 a, b.

18. 线性模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, $\mathrm{Var}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{v}$. 为了 $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ 为 σ^2 的无偏估计，$\boldsymbol{A}$ 应满足什么条件？

19. 设观察值 $y_1, \cdots, y_n$ 与参数 $\beta_1, \cdots, \beta_q$ 的正确关系是线性模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}=(\boldsymbol{X}_p \vdots \boldsymbol{X}_R)\begin{pmatrix}\boldsymbol{\beta}_p \\ \boldsymbol{\beta}_R\end{pmatrix}+\boldsymbol{\varepsilon}$, 而我们由于不了解，或为计算简便计，采用了模型 $\boldsymbol{y}=\boldsymbol{X}_p\boldsymbol{\beta}_p+\boldsymbol{\varepsilon}$. 设 $(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}$ 存在. 设我们从后一模型中算出了 $\boldsymbol{\beta}_p$ 的 LS 估计 $\widetilde{\beta}_p$. 问 $\widetilde{\beta}_p$ 为 $\boldsymbol{\beta}_p$ 的无偏估计的充分必要条件是什么？解释这些条件的意义.

20. 试证对模型(6.53)附加的约束条件(6.54)满足(6.74)和(6.76).

第七章 多元分析基础

在本章中将集中讨论多指标的统计分析问题．按照习惯称这一分支为多元统计分析，简称为多元分析．

在实际问题中，我们往往对对象的多个指标感兴趣．例如：在天气预报中，要同时预报气温、气压、风向、风力及降水等多项指标；在检查人体的健康状况时，需要检查体温、血压、肺活量、脉搏以及血液等各方面情况；在分析矿石标本时，我们将关心它所含的种种元素和化合物．总之，在这类问题中，我们要观察取自总体的个体的多个指标值，得到多维的样本观察值，在此基础上进行统计分析．

在上一章讨论回归分析的时候，虽然也涉及多个随机变量的情形，但那时我们将某个随机变量看成因变量，而将其余视作自变量，突出了对一个指标的关心．在本章的多元回归分析中，我们将把多个随机变量当作因变量，讨论一组随机变量对另一组随机变量的依赖性．从而可见这里所说的多元统计分析与一元统计分析的分野．当然，在某些情况下，这种区分并不是很严格的，也可能并不重要．

在实际生活中需要关心多指标的情形，明显地更为常见．尤其是在地质、气象、生物、经济、军事等领域，免不了要和许多指标打交道．因而多元分析有十分广泛的应用价值．

近二十多年来，多元分析发展迅速，普及甚广，其推动力固然是实际需要，也离不开计算机的飞跃发展及普遍使用．在统计分析中，随着指标的增加，不但问题变得更为复杂，而且计算量将大大增加，况且统计分析常常有时间上的紧迫性(如天气或地震的预报)，如果没有高速、大容量的计算机，很难想像能及时作出分析结

果. 为了提高计算的效率，多元统计分析将更多地依赖计算机算法.

另一个值得注意的情况，正如 M. Kendell 在《多元分析》一书的序言中所说"因为多元分析这门学科正在发展之中，它虽然是一门科学，但也可看作是一种艺术。"这段话意味着多元分析方法的不确切性和灵活性. 它可以从两个方面来理解：一方面，多元分析理论还很不成熟，尽管从各个角度提出了不少可以使用的统计方法，但这些方法尚缺乏充分的理论依据，对所得结果也往往难以作出较完善的统计解释；另一方面，由于多元统计问题的复杂性和不确定性，要从理论上给出严谨系统的方法，至少在当前还很难做到.

作为数理统计基础教程中的一章，远远不能容纳多元分析的丰富内容. 我们的目标是给出多元分析的最基础的理论，并讨论一些常用的在理论上较有根据的方法，使读者能从中窥见多元分析之一斑. 至于多元分析中很重要的算法问题，就唯有忍痛割爱了.

§7.1 多元正态总体的抽样分布及参数推断

如同在一元统计分析中相仿，正态分布在多元统计分析中也起着基本作用. 由于推导多元统计量的分布困难更大，在多元分析理论中，凡涉及到需要给出统计量分布的场合，几乎都是在假定总体有多元正态分布的基础上展开的. 因此，本节所介绍的理论，无疑是多元分析的必要基础.

(一) 多元正态分布及其基本性质

我们所熟知的正态随机变量 X，一般有密度函数为

$$f(x)=(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \qquad (7.1)$$

其中 μ, σ^2 分别为 X 的均值和方差. 记 $X\sim N_1(\mu,\ \sigma^2)$. 正态

随机变量的分布由它的均值和方差完全决定.

如果 $X_1, \cdots, X_p$ 是来自正态总体 $N_1(\mu, \sigma^2)$ 的随机样本,那么 $\boldsymbol{X}=(X_1, \cdots, X_p)'$ 的密度函数是

$$f(x_1, \cdots, x_n)=(2\pi\sigma^2)^{-\frac{p}{2}}\exp\left\{-\frac{1}{2\sigma^2}\Sigma_1^p(x_i-\mu)^2\right\}. \tag{7.2}$$

这里的 p 维随机向量 X 被认为有 p 元正态分布. 但是,值得注意的是一般的多元正态分布,并不总有密度函数,有形如(7.2)的密度函数的,只是其很特殊的情形. 为此,我们给多元正态分布下一个确切的定义,然后讨论它的一些基本性质.

定义 7.1

设 $\boldsymbol{X}=(X_1,\cdots,X_p)'$ 是 p 维随机向量. 如果对任一 p 维实向量 $\boldsymbol{a}\in R^p$, 都有 $\boldsymbol{a}'\boldsymbol{X}$ 是正态随机变量,则称 $\boldsymbol{X}$ 是 **$\boldsymbol{p}$ 维正态(随机)向量**,称 $\boldsymbol{X}$ 的分布为 **$\boldsymbol{p}$ 元正态分布**.

记 $\boldsymbol{X}$ 的均值向量 $E\boldsymbol{X}$ 为 $\boldsymbol{\mu}$, 记 $\boldsymbol{X}$ 的协方差矩阵 $\mathrm{Cov}\boldsymbol{X}$ 为 $\boldsymbol{\Sigma}$. 下面的附注将说明 $\boldsymbol{X}$ 的分布被 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 唯一决定, 因而可记 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 注意我们并没有假定 $\boldsymbol{\Sigma}$ 是可逆阵.

附注 7.1 p 元正态分布可以用许多种等价的方法来定义,这些等价条件都称作正态分布的刻划(Charaterization). 其中最重要的一种就是用特征函数. 一般概率论教科书中都给出了随机变量 $\boldsymbol{X}$ 的特征函数的表示式为 $Ee^{it\boldsymbol{X}}$. 正态随机变量 $X\sim N_1(\mu, \sigma^2)$ 的特征函数是 $\exp\left\{i\mu t-\frac{1}{2}\sigma^2t^2\right\}$. **随机向量 $\boldsymbol{X}$ 的特征函数**定义为

$$\phi_X(\boldsymbol{t})=Ee^{i\boldsymbol{t}'\boldsymbol{X}}. \tag{7.3}$$

它是 $\boldsymbol{t}=(t_1, \cdots, t_p)'$ 的函数. 根据特征函数的逆转公式, $\phi_X(\boldsymbol{t})$ 唯一地决定 $\boldsymbol{X}$ 的分布函数. 设 $\boldsymbol{X}$ 是正态随机向量, $\boldsymbol{X}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 对于 $\boldsymbol{t}\in R^p$, 有 $\boldsymbol{t}'\boldsymbol{X}$ 是正态随机变量,它的特征函数应当是

$$\phi(\theta)=Ee^{i\theta\boldsymbol{t}'\boldsymbol{X}}=\exp\left\{i\theta E\boldsymbol{t}'\boldsymbol{X}-\frac{1}{2}D(\boldsymbol{t}'\boldsymbol{X})\theta^2\right\}.$$

注意到 $E\boldsymbol{t}'\boldsymbol{X}=\boldsymbol{t}'E\boldsymbol{X}=\boldsymbol{t}'\boldsymbol{\mu},\ D(\boldsymbol{t}'\boldsymbol{X})=\boldsymbol{t}'\mathrm{Cov}\boldsymbol{X}\boldsymbol{t}=\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t},$

于是在 $\phi(\theta)$ 的表达式中取 $\theta=1$ 就得到

$$Ee^{it'X}=\exp\left\{it'\mu-\frac{1}{2}t'\Sigma t\right\}. \tag{7.4}$$

这正是 X 的特征函数. 它由 μ 和 Σ 所决定, 故可断言正态向量的均值 μ 和协方差阵 Σ 唯一地决定了它的分布.

如果读者缺乏特征函数知识, 可以只接受附注的结论而不顾其理论推导.

下面是多元正态分布的基本性质:

性质1 正态随机向量的线性函数必定是正态的. 设 $X\sim N_p(\mu, \Sigma)$; A 是任一 $q\times p$ 阶的实矩阵, b 是 q 维实向量, 那么

$$Y\triangleq AX+b\sim N_q(A\mu+b,\ A\Sigma A'). \tag{7.5}$$

证 任给 $a\in R^q$, 有 $a'Y=a'AX+a'b$, 因 $a'AX$ 是正态变量, $a'b$ 是常数, 知 $a'Y$ 是正态变量, 因此 Y 是正态向量. 其余显然.

性质2 对于任给的 p 维实向量 μ 和 p 阶非负定方阵 Σ, 必存在正态随机向量 X 以 μ 为均值向量, 以 Σ 为协方差矩阵.

证 由第六章附录, 知 Σ 有分解式 $\Sigma=AA'$. 取 p 个相互独立的标准正态随机变量 $Y_1,\cdots,Y_p$, 记 $Y=(Y_1,\ \cdots,\ Y_p)'$, 则有 $Y\sim N_p(0,\ I_p)$. (看(7.2))令 $X=AY+\mu$, 由性质1立得

$$X\sim N_p(\mu,\ AA')=N_p(\mu,\ \Sigma).$$

证毕.

这一证明实际上蕴涵了下面的

性质3 任一 p 维正态随机向量, 可表为 p 维标准正态向量 $(\sim N_p(0,\ I_p))$ 的线性函数.

性质4 设 $X\sim N_p(\mu,\ \Sigma)$. 如果 Σ 正定(因 Σ 原是非负定阵, 只需要求其可逆就够了), 称 X 是非退化的. 此时, X 有密度函数

$$f(x)=(2\pi)^{-\frac{p}{2}}(\det\Sigma)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}. \tag{7.6}$$

证 用性质3给出 $X=\Sigma^{\frac{1}{2}}Y+\mu,\ Y\sim N_p(0,\ I_p)$. 由(7.2)知 Y 的密度函数为

$$f(\boldsymbol{y})=(2\pi)^{-\frac{p}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{y}'\boldsymbol{y}\right\}$$

作变换$\boldsymbol{y}=\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})$，知 Jacobi 行列式是 $\det\boldsymbol{\Sigma}^{-\frac{1}{2}}$，因此$f(\boldsymbol{X})$的密度函数如(7.6)。

从性质 4 的证明可以看出，如果 $\boldsymbol{\Sigma}$ 奇异，那么变换$\boldsymbol{x}=\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{y}+\boldsymbol{\mu}$ 不是一一的，积分变量替换将无法进行．这时 $\boldsymbol{X}$ 的分布函数是

$$F(\boldsymbol{x})=\int_{\Sigma^{\frac{1}{2}}y<x-\mu}f(\boldsymbol{y})d\boldsymbol{y}, \tag{7.7}$$

但不再存在密度函数，其中 $\boldsymbol{a}<\boldsymbol{b}$ 表示 $a_i<b_i$, $i=1, \cdots, p$.

性质 5 设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$．相应地作如下剖分

$$\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_{(1)}\\ \boldsymbol{X}_{(2)}\end{pmatrix},\ \boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_{(1)}\\ \boldsymbol{\mu}_{(2)}\end{pmatrix},\ \boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix},$$

其中 $\boldsymbol{X}_{(i)}$ 是 p_i 维的，$i=1, 2$, $p_1+p_2=p$．则有

1° 如果 $\boldsymbol{\Sigma}_{12}=\boldsymbol{\Sigma}'_{21}=0$，则有 $\boldsymbol{X}_{(1)}$与 $\boldsymbol{X}_{(2)}$独立；

2° 如果 $\boldsymbol{\Sigma}_{11}$ 可逆，则在给定 $\boldsymbol{X}_{(1)}$ 时 $\boldsymbol{X}_{(2)}$的条件分布是

$$N_{p_2}(\boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_{(1)}-\boldsymbol{\mu}_{(1)}),\ \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \tag{7.8}$$

当 $\boldsymbol{\Sigma}_{22}$ 可逆，有类似的结论．

证 1° 用性质 4 证明中的方法，从(7.7)式不难给出证明(用特征函数更易看出)，留给读者作为练习．

2° 的证明如下：令

$$\boldsymbol{P}=\begin{pmatrix}\boldsymbol{I}_{p_1} & \boldsymbol{0}\\ -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{I}_{p_2}\end{pmatrix}$$

作变换 $\boldsymbol{Y}=\boldsymbol{P}\boldsymbol{X}$，则有

$$\boldsymbol{Y}=\begin{pmatrix}\boldsymbol{X}_{(1)}\\ \boldsymbol{X}_{(2)}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_{(1)}\end{pmatrix}\triangleq\begin{pmatrix}\boldsymbol{Y}_{(1)}\\ \boldsymbol{Y}_{(2)}\end{pmatrix},$$

$$\operatorname{Cov}\boldsymbol{Y}=\boldsymbol{P}\operatorname{Cov}\boldsymbol{X}\boldsymbol{P}'=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\end{pmatrix}.$$

知 $\boldsymbol{Y}_{(1)}$ 与 $\boldsymbol{Y}_{(2)}$ 独立．因此 $\boldsymbol{Y}_{(1)}(=\boldsymbol{X}_{(1)})$给定时，$\boldsymbol{Y}_{(2)}$ 的条件分布

也就是 $\boldsymbol{Y}_{(2)}$ 的边缘分布，故有

$$\boldsymbol{Y}_{(2)} \sim N_{p_2}(\boldsymbol{\mu}_{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_{(1)},\ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}).$$

但 $\boldsymbol{X}_{(2)} = \boldsymbol{Y}_{(2)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_{(1)}$，得 $\boldsymbol{X}_{(1)}$ 给定时 $\boldsymbol{X}_{(2)}$ 的条件分布如(7.8)式. 证毕.

性质5.1°表明不相关的正态随机向量 $\boldsymbol{X}$, $\boldsymbol{Y}$,（即满足 $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = 0$）一定是独立的. 反之当然也成立. 2°的特殊情形 $p_1 = p-1$, $p_2 = 1$ 在多元分析中很有用. 从(7.8)可见 $\boldsymbol{X}_{(1)}$ 给定时 $\boldsymbol{X}_{(2)}$ 的条件均值是 $\boldsymbol{X}_{(1)}$ 的线性函数，这正是讨论一元统计时对正态总体的回归所得结论的推广，在多元回归中亦将起重要作用.

（二）正态总体的抽样分布

设总体 G 有 p 元正态分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ 是取自 G 的简单随机样本，即有

$$\boldsymbol{x}_\alpha \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \mathrm{Cov}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) = 0,\ \beta, \alpha = 1, \cdots, n,\ \beta \neq \alpha.$$

记

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix},$$

我们来研究随机矩阵 $\boldsymbol{X}$ 的分布. 记 $\boldsymbol{x}' = (\boldsymbol{x}_1' \cdots \boldsymbol{x}_n')$，易见 $E\boldsymbol{x}' = (\boldsymbol{\mu}' \cdots \boldsymbol{\mu}')$，且有

$$\mathrm{Cov}\boldsymbol{x} = \begin{pmatrix} \boldsymbol{\Sigma} & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{\Sigma} \end{pmatrix}.$$

可见 $\boldsymbol{x}$ 遵从 np 元正态分布. 设 $\boldsymbol{\Sigma} = \boldsymbol{B}'\boldsymbol{B}$，由（一）性质2，存在由相互独立的标准正态变量 $z_{ij}(\sim N(0, 1))$ 为元素的矩阵 $\boldsymbol{Z}$，使得 $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{B} + \boldsymbol{M}$, $\boldsymbol{M} = E\boldsymbol{X}$.

将上述情形推广到 $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z}\boldsymbol{B} + \boldsymbol{M}$. 这里 $\boldsymbol{Z} = (z_{ij})$ 是 $n \times p$ 阶随机阵，z_{ij} 相互独立遵从 $N(0, 1)$，$\boldsymbol{A}$ 是 $m \times n$ 阶实数阵，$\boldsymbol{B}$ 是 $p \times q$ 阶实数阵，$\boldsymbol{M}$ 是 $m \times q$ 阶实数阵，且记 $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{V} = (v_{ij})$，$\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{\Sigma} = (\sigma_{ij})$. 显然，将 $\boldsymbol{Y}$ 看作 mq 维向量，有 $\boldsymbol{Y}$ 遵从 mq 元正态分布. 计算可得

$$Ey_{ij} = m_{ij},$$

$$D(y_{ij})=D(\sum_{\alpha,\beta} a_{i\alpha}z_{\alpha\beta}b_{\beta j})=\sum_{\alpha,\beta} a_{i\alpha}^2 b_{\beta j}^2=\sum_{\alpha} a_{i\alpha}^2\cdot\sum_{\beta} b_{\beta j}^2=v_{ii}\sigma_{jj},$$

$$\begin{aligned}\mathrm{Cov}(y_{ij},\ y_{kh})&=\mathrm{Cov}(\sum_{\alpha,\beta} a_{i\alpha}z_{\alpha\beta}b_{\beta j},\ \sum_{\alpha,\beta} a_{k\alpha}z_{\alpha\beta}b_{\beta h})\\&=\sum_{\alpha,\beta} a_{i\alpha}b_{\beta j}a_{k\alpha}b_{\beta h}\\&=v_{ik}\sigma_{jh},\quad i,\ k=1,\ \cdots,\ m;\ h,\ j=1,\ \cdots,\ q.\end{aligned}$$

因此 $\boldsymbol{Y}$ 的一、二阶矩仅依赖 $\boldsymbol{M}$, $\boldsymbol{V}$ 和 $\boldsymbol{\Sigma}$. 记

$$\boldsymbol{Y}\sim N_{mq}(\boldsymbol{M},\ \boldsymbol{V},\ \boldsymbol{\Sigma}).\tag{7.9}$$

于是有前面的 $\boldsymbol{X}\sim N_{np}(\boldsymbol{M},\ \boldsymbol{I},\ \boldsymbol{\Sigma})$. 在多元统计分析中,我们仅限于讨论如(7.9)的矩阵正态分布.

定理 7.1

设 $\boldsymbol{Y}\sim N_{mq}(\boldsymbol{M},\ \boldsymbol{V},\ \boldsymbol{\Sigma})$. 任给 $\boldsymbol{C}$, $\boldsymbol{D}$ 分别为 $k\times m$ 阶和 $q\times r$ 阶实阵,则有

$$\boldsymbol{X}=\boldsymbol{CYD}\sim N_{kr}(\boldsymbol{CMD},\ \boldsymbol{CVC'},\ \boldsymbol{D'\Sigma D}).$$

证明是容易的,留作练习.

作为一元统计中 χ^2 变量的推广,我们研究

$$\boldsymbol{W}=\boldsymbol{X'X},\ \boldsymbol{X}\sim N_{np}(\boldsymbol{M},\ \boldsymbol{I},\ \boldsymbol{\Sigma}),\ \boldsymbol{M'}=(\boldsymbol{\mu}_1\cdots\boldsymbol{\mu}_n)$$

的分布. 这类随机矩阵的分布,习惯上称 Wishart 分布. 它的密度函数问题比较复杂,本书不予讨论,但有必要研究它的一些重要性质.

我们不难验算

$$E\boldsymbol{W}=\sum_{\alpha=1}^{n} E\boldsymbol{x}_\alpha\boldsymbol{x}'_\alpha=n\boldsymbol{\Sigma}+\boldsymbol{M'M}.$$

其中 $\boldsymbol{M'}=E\boldsymbol{X'}=(\boldsymbol{\mu}_1\cdots\boldsymbol{\mu}_n)$.

记 $\boldsymbol{\Delta}=\boldsymbol{M'M}$, 通过不在此给出的推理,可见一个遵从 Wishart 分布的随机矩阵,从分布的角度而言,仅仅依赖于 p, n, $\boldsymbol{\Sigma}$ 和 $\boldsymbol{\Delta}$. 称 p 是 W 的维数, n 是自由度, $\boldsymbol{\Sigma}$ 是所依赖的协方差矩阵, $\boldsymbol{\Delta}$ 是非中心矩阵, 记 W 的分布为

$$W_p(n,\ \boldsymbol{\Sigma},\ \boldsymbol{\Delta})$$

全称为 p 维的基于协方差矩阵 $\boldsymbol{\Sigma}$ 的自由度为 n、非中心矩阵为 $\boldsymbol{\Delta}$ 的 **Wishart 分布**. 简称为 Wishart 分布. 依 $\boldsymbol{\Delta}$ 是否为零而称

中心 Wishart 分布和非中心 Wishart 分布．如果一个随机矩阵可表成(7.9)的形式，就认为它遵从 $W_p(n, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$ 分布．（注意此分布仅依赖 $\boldsymbol{\Delta}=\boldsymbol{M}'\boldsymbol{M}$.）

定理 7.2

设 $W\sim W_p(n, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$，则有

1° 如果 $p=1$, $\Sigma=\sigma^2$, $W\sim\sigma^2\chi_n^2(\Delta^{\frac{1}{2}}/\sigma)$;

2° 设 $\boldsymbol{A}$ 是任意的 $k\times p$ 阶阵，有

$$\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}'\sim W_k(n, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}', \boldsymbol{A}\boldsymbol{\Delta}\boldsymbol{A}').$$

证 1° 是显然的.

2° 记 $\boldsymbol{Y}'=\boldsymbol{A}\boldsymbol{X}'=(\boldsymbol{y}_1\cdots\boldsymbol{y}_n)$，易见

$\boldsymbol{y}_\alpha\sim N_k(\boldsymbol{A}\boldsymbol{\mu}_\alpha, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$ 且相互独立，$\alpha=1, \cdots, n$. 因此 $\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}'=\boldsymbol{Y}'\boldsymbol{Y}\sim W_k(n, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}', \boldsymbol{A}\boldsymbol{\Delta}\boldsymbol{A}')$. 得证.

定理 7.3

设 $\boldsymbol{X}\sim N_{np}(\boldsymbol{M}, \boldsymbol{I}, \boldsymbol{\Sigma})$, $\boldsymbol{C}=\boldsymbol{C}'$，则有

1° $\boldsymbol{W}\triangleq\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}\sim W_p(k, \boldsymbol{\Sigma}, \boldsymbol{M}'\boldsymbol{C}\boldsymbol{M})\Leftrightarrow\boldsymbol{C}$ 为正投影阵，$rk\boldsymbol{C}=k$;

2° $\boldsymbol{A}\boldsymbol{X}$ 与 $\boldsymbol{B}\boldsymbol{X}$ 独立 $\Leftrightarrow\boldsymbol{A}\boldsymbol{B}'=0$; 从而有：$\boldsymbol{B}\geqslant\boldsymbol{O}$, $\boldsymbol{A}\boldsymbol{B}=\boldsymbol{O}$ 推出 $\boldsymbol{A}\boldsymbol{X}$ 与 $\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}$ 独立，$\boldsymbol{A}\geqslant0$, $\boldsymbol{B}\geqslant0$, $\boldsymbol{A}\boldsymbol{B}=0$ 推出 $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ 与 $\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}$ 独立.

证 1° "$\Rightarrow$" 对一切 $\boldsymbol{a}\in R^p$. 有 $\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}\sim\chi_k^2(\delta)$，得 $\boldsymbol{C}$ 为正投影阵. "$\Leftarrow$" 存在 $\underset{n\times k}{\boldsymbol{U}}$ 满足 $\boldsymbol{C}=\boldsymbol{U}\boldsymbol{U}'$, $\boldsymbol{U}'\boldsymbol{U}=\boldsymbol{I}_k$，令 $\boldsymbol{Y}=\boldsymbol{U}'\boldsymbol{X}$, 记 $\boldsymbol{U}=(\boldsymbol{u}_1\cdots\boldsymbol{u}_k)$ 有

$$\operatorname{Cov}(\boldsymbol{X}'\boldsymbol{u}_i, \boldsymbol{X}'\boldsymbol{u}_j)=\operatorname{Cov}(\sum_\alpha\boldsymbol{x}_\alpha u_{i\alpha}, \sum_\alpha\boldsymbol{x}_\alpha u_{\alpha j})$$

$$=\delta_{ij}\boldsymbol{\Sigma}=\begin{cases}\boldsymbol{\Sigma}, & i=j,\\ 0, & i\neq j.\end{cases}$$

于是 $\boldsymbol{X}'\boldsymbol{C}\boldsymbol{X}=\boldsymbol{Y}'\boldsymbol{Y}$，而 $\boldsymbol{Y}\sim N_{kp}(\boldsymbol{U}'\boldsymbol{M}, \boldsymbol{I}, \boldsymbol{\Sigma})$，得证.

2° 的证明可由对一切 i, j

$\operatorname{Cov}(\Sigma\boldsymbol{x}_\alpha a_{i\alpha}, \Sigma\boldsymbol{x}_\alpha b_{j\alpha})=\sum_\alpha a_{i\alpha}b_{j\alpha}\cdot\boldsymbol{\Sigma}=\boldsymbol{0}\Leftrightarrow\boldsymbol{A}\boldsymbol{B}'=\boldsymbol{0}$，得到.

定理 7.4

设 $W\sim W_p(n, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$, $n\geqslant p$, $\boldsymbol{\Sigma}>\boldsymbol{0}$. 则有

$$P(\boldsymbol{W}>\boldsymbol{0})=1. \tag{7.10}$$

证 按定义有 $\boldsymbol{W}=\boldsymbol{X}'\boldsymbol{X}$, $\boldsymbol{X}\sim N_{np}(\boldsymbol{M}, \boldsymbol{I}, \boldsymbol{\Sigma})$, $\boldsymbol{M}'\boldsymbol{M}=\boldsymbol{\Delta}$, 记 $\boldsymbol{X}=(\boldsymbol{x}_{(1)}|\cdots|\boldsymbol{x}_{(p)})$, 对 $k<p$, 由正态分布性质5得: 当 $\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(k)}$ 给定时, $\boldsymbol{x}_{(k+1)}$ 的条件分布是 $N(\boldsymbol{\nu}_{k+1}, \sigma_{k+1}^2\boldsymbol{I})$, 这里 $\sigma_{k+1}^2=\boldsymbol{e}_1'(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})\boldsymbol{e}_1$, 其中 $\boldsymbol{e}_1'=(10\cdots0)$. 因此条件分布非退化, 故有

$$P(\boldsymbol{x}_{(k+1)}\in\mu(\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(k)})\,|\,\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(k)})=0.$$

这是由于 $\mu(\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(k)})$ 至多是 k 维子空间, 而 $\boldsymbol{x}_{(k+1)}$ 是 n 维的, 且条件分布非退化. 从而有

$$\begin{aligned}&P(\boldsymbol{x}_{(k+1)}\in\mu(\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(k)}))\\&\quad=E\{P(\boldsymbol{x}_{(k+1)}\in\mu(\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(k)})\,|\,\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(k)})\}=0.\end{aligned}$$

于是 $\quad P(rk\boldsymbol{X}<p)=P(\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(p)}\text{ 线性相关})$

$$\leqslant\sum_{k=0}^{p-1}P(\boldsymbol{x}_{(k+1)}\in\mu(\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(k)}))=0.$$

因此有 $P(\boldsymbol{X}'\boldsymbol{X}>0)=1$, 得(7.11).

根据定理7.4, 当 $n\geqslant p$, $\boldsymbol{\Sigma}>\boldsymbol{0}$, 只要忽略一个概率为零的事件, 就可认为 $\boldsymbol{W}$ 是可逆的, 称 $\boldsymbol{W}$ 有非退化 Wishart 分布. 当 $n<p$ 或 $\det\boldsymbol{\Sigma}=0$, 称 $\boldsymbol{W}$ 有退化的 Wishart 分布.

作为 Cochran 定理到"矩阵二次型"的推广, 有

定理 7.5

设 $\boldsymbol{X}\sim N_{np}(\boldsymbol{M}, \boldsymbol{I}, \boldsymbol{\Sigma})$, $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_n\geqslant 0$, 满足

$$\boldsymbol{X}'\boldsymbol{X}=\sum_{i=1}^{k}\boldsymbol{X}'\boldsymbol{A}_i\boldsymbol{X}, \text{记 } rk\boldsymbol{A}_i=n_i,$$

则 $\boldsymbol{X}'\boldsymbol{A}_i\boldsymbol{X}\sim W_p(n_i, \boldsymbol{\Sigma}, \boldsymbol{M}'\boldsymbol{A}_i\boldsymbol{M})$, $i=1, \cdots, k\Leftrightarrow n=\sum_{i=1}^{k}n_i$.

证明可从 $\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}$ 推出. 留作习题.

引理 7.1

设 $W\sim W_p(n, \boldsymbol{\Sigma})$, $n\geqslant p$, $\boldsymbol{\Sigma}>0$. 将 $\boldsymbol{W}$ 和 $\boldsymbol{\Sigma}$ 相应地剖分为

$$\boldsymbol{W}=\begin{pmatrix}\boldsymbol{W}_{11} & \boldsymbol{W}_{12}\\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22}\end{pmatrix}, \boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$$

其中 $\boldsymbol{W}_{11}, \boldsymbol{\Sigma}_{11}$ 为 q 阶方阵. 记 $\boldsymbol{W}_{11.2}=\boldsymbol{W}_{11}-\boldsymbol{W}_{12}\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21}$, $\boldsymbol{\Sigma}_{11.2}=\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. 则有

$\boldsymbol{W}_{22} \sim W_{p-q}(n, \boldsymbol{\Sigma}_{22})$，在 $\boldsymbol{W}_{22}$ 给定的条件下

$(\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21} \mid \boldsymbol{W}_{22}) \sim N_{p-q,q}(\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \boldsymbol{W}_{22}^{-1}, \boldsymbol{\Sigma}_{11.2})$，

$(\boldsymbol{W}_{11.2} \mid (\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21}, \boldsymbol{W}_{22})) \sim \boldsymbol{W}_q(n-p+q, \boldsymbol{\Sigma}_{11.2})$，不依赖$(\boldsymbol{W}_{21}, \boldsymbol{W}_{22})$.

证 在定理7.2.2°中，取

$$\boldsymbol{A} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-q} \end{pmatrix}, \quad \text{立得 } \boldsymbol{W}_{22} \sim \boldsymbol{W}_{p-q}(n, \boldsymbol{\Sigma}_{22}).$$

设 $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$, $\boldsymbol{X} \sim N_{np}(\mathbf{0}, \boldsymbol{I}, \boldsymbol{\Sigma})$，记 $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$，这里 $\boldsymbol{X}_1$ 有 q 列. 因 $\boldsymbol{X}$ 的行独立同分布，由正态分布性质5.2°，不难得到在 $\boldsymbol{X}_2$ 给定时 $\boldsymbol{X}_1$ 的条件分布为

$$\boldsymbol{X}_1 \mid \boldsymbol{X}_2 \sim N_{nq}(\boldsymbol{X}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \boldsymbol{I}, \boldsymbol{\Sigma}_{11.2}) \qquad (7.11)$$

注意到 $\boldsymbol{W}_{11} = \boldsymbol{X}_1'\boldsymbol{X}_1$, $\boldsymbol{W}_{21} = \boldsymbol{X}_2'\boldsymbol{X}_1$, $\boldsymbol{W}_{22} = \boldsymbol{X}_2'\boldsymbol{X}_2$，由定理7.1得

$$\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21} = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1 \mid \boldsymbol{X}_2 \sim N_{p-q,q}(\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}, \boldsymbol{\Sigma}_{11.2}).$$

知它仅依赖于 $\boldsymbol{X}_2'\boldsymbol{X}_2 = \boldsymbol{W}_{22}$，故有 $\boldsymbol{W}_{22}$ 给定时

$$\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21} \mid \boldsymbol{W}_{22} \sim N_{p-q,q}(\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \boldsymbol{W}_{22}^{-1}, \boldsymbol{\Sigma}_{11.2}). \qquad (7.12)$$

由于 $\boldsymbol{W}_{11.2} = \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2')\boldsymbol{X}_1 = \boldsymbol{X}_1'\boldsymbol{P}_{\boldsymbol{X}_2^\perp}\boldsymbol{X}_1$，

由定理7.3.1°，当 $\boldsymbol{X}_2$ 给定时有

$$\boldsymbol{W}_{11.2} \mid \boldsymbol{X}_2 \sim W_q(n-p+q, \boldsymbol{\Sigma}_{11.2})$$

(非中心矩阵$\boldsymbol{\Delta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{X}_2^\perp}\boldsymbol{X}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = 0$). 又由定理7.3.2°，由$(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{X}_2^\perp} = 0$ 得 $\boldsymbol{W}_{11.2}$ 与 $\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21}$ 在 $\boldsymbol{X}_2$ 给定的条件下独立，而 $\boldsymbol{W}_{11.2}$ 与 $\boldsymbol{X}_2$ 无关，$\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21}$ 仅依赖 $\boldsymbol{W}_{22}$，故有

$$\boldsymbol{W}_{11.2} \mid (\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21}, \boldsymbol{W}_{22}) \sim W_q(n-p+q, \boldsymbol{\Sigma}_{11.2}). \qquad \text{证毕.}$$

定理7.6

设 $\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma})$, $n \geqslant p$, $\boldsymbol{\Sigma} > 0$，则有

1° 对任给的 $\boldsymbol{a} \in R^p$，有

$$\frac{\boldsymbol{a}'\boldsymbol{\Sigma}^{-1}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}^{-1}\boldsymbol{a}} \sim \chi^2_{n-p+1};$$

2° 设 $\boldsymbol{x} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$，$\boldsymbol{x}$ 与 $\boldsymbol{W}$ 独立，有

$$x'W^{-1}x\cdot\frac{n-p+1}{p}\sim F_{p,n-p+1}.\qquad(7.13)$$

证 先考虑 $a=e_1=(10\cdots0)'$，有

$$\frac{a'\Sigma^{-1}a}{a'W^{-1}a}=\frac{(\Sigma^{-1})_{11}}{(W^{-1})_{11}}=\frac{(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}}{(W_{11}-W_{12}W_{22}^{-1}W_{21})^{-1}}$$

$$=\frac{W_{11}-W_{12}W_{22}^{-1}W_{21}}{\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$$

用引理7.1 $q=1$ 的情形得 $\frac{e_1'\Sigma^{-1}e_1}{e_1'W^{-1}e_1}\sim\chi^2_{n-p+1}$. 对 $a\in R^p$, 令 $u_1=a/\|a\|$，作正交阵 $U=(u_1,\ \cdots u_p)$，有 $a=Ue_1\|a\|$，得

$$\frac{a'\Sigma^{-1}a}{a'W^{-1}a}=\frac{e_1'U'\Sigma^{-1}Ue_1}{e_1'U'W^{-1}Ue_1}=\frac{e_1'(U'\Sigma U)^{-1}e_1}{e_1'(U'WU)^{-1}e_1}.$$

由定理 7.2 知 $U'WU\sim W_p(n,\ U'\Sigma U)$, 故由已得结论立得 $\frac{a'\Sigma^{-1}a}{a'W^{-1}a}\sim\chi^2_{n-p+1}$ 对一切 $a\in R^p$ 成立.

注意到
$$x'W^{-1}x=\frac{x'\Sigma^{-1}x}{x'\Sigma^{-1}x/x'W^{-1}x},$$

当 x 给定时，有$\left(\frac{x'\Sigma^{-1}x}{x'W^{-1}x}\,\middle|\,x\right)\sim\chi^2_{n-p+1}$. 因此条件分布实际上不依赖于 x，得 $\frac{x'\Sigma^{-1}x}{x'W^{-1}x}\sim\chi^2_{n-p+1}$, 且知它与 x 独立，从而也与 $x'\Sigma^{-1}x$ 独立. 于是有(7.13)成立. 证毕.

记 $T^2=x'W^{-1}x$，称它为 **Hotelling T^2 统计量**，它在多元统计分析的假设检验中起着重要作用.

设 $W_1\sim W_p(k_1,\ \Sigma)$, $W_2\sim W_p(k_2,\ \Sigma)$，且 W_1 与 W_2 独立，令

$$\Lambda=\frac{\det W_1}{\det(W_1+W_2)}.\qquad(7.14)$$

称 Λ 为 Wilks 统计量，它的分布仅仅依赖于 p, k_1 和 k_2，记它的分布为 $\Lambda(p,\ k_1,\ k_2)$. Wilks Λ 统计量在假设检验中也很重要. 但它的精确分布的推导相当困难，直到 1966 年才由 Schatzoff 求出. 当 $p=1, 2$, 或 $k_2=1,\ 2$ 时，$\Lambda(p,\ k_1,\ k_2)$与 F 分布有密切联系，其联系见表 7.1.

一般情形的 Λ 分布表($p\leqslant8$)可看张尧庭、方开泰《多元统计

分析引论》.

表7.1　$\Lambda(p, k_1, k_2)$与F分布$(k_1>p)$

p	k_2	F 分 布	自 由 度
	1	$\frac{1-\Lambda}{\Lambda}\cdot\frac{k_1-p+1}{p}$	$p,\ k_1-p+1$
	2	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\cdot\frac{k_1-p}{p}$	$p,\ k_1-p$
1		$\frac{1-\Lambda}{\Lambda}\cdot\frac{k_1}{k_2}$	$k_2,\ k_1$
2		$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\cdot\frac{k_1-1}{k_2}$	$2k_2,\ 2(k_1-1)$

(三)正态总体的参数统计推断

设总体G有p元正态分布$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$是取自$G$的容量为$n$的简单随机样本,记

$$\boldsymbol{Y}'=(\boldsymbol{y}_1\cdots\boldsymbol{y}_n),\text{有 } \boldsymbol{Y}\sim N_{np}(\boldsymbol{M}, \boldsymbol{I}, \boldsymbol{\Sigma})$$

这里$\boldsymbol{M}=(\boldsymbol{\mu}\cdots\boldsymbol{\mu})'$.

仿一元统计,令样本均值和样本协方差阵为

$$\bar{\boldsymbol{y}}=\frac{1}{n}\sum_1^n\boldsymbol{y}_\alpha,\quad S=\frac{1}{n-1}\sum_1^n(\boldsymbol{y}_\alpha-\bar{\boldsymbol{y}})(\boldsymbol{y}_\alpha-\bar{\boldsymbol{y}})'.$$

用矩阵记法有

$$\bar{\boldsymbol{y}}=\boldsymbol{Y}'\boldsymbol{1}_n\frac{1}{n},\ S=\boldsymbol{Y}'\boldsymbol{P}_{1_n^\perp}\boldsymbol{Y}\frac{1}{n-1}.$$

先讨论$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$的估计. 在上述记号下有

定理 7.7

$\bar{\boldsymbol{y}}$和$\boldsymbol{S}$分别是$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$的无偏估计. 并且,在$\boldsymbol{\mu}$的一切线性无偏估计$\{\boldsymbol{Y}'\boldsymbol{a}\}$中,$\bar{\boldsymbol{y}}$满足

$$\mathrm{Cov}(\boldsymbol{Y}'\boldsymbol{a})-\mathrm{Cov}\bar{\boldsymbol{y}}\geqslant 0.\ (\text{非负定}) \tag{7.15}$$

证　容易验算

$$E\bar{\boldsymbol{y}}=E\boldsymbol{Y}'\boldsymbol{1}_n\frac{1}{n}=\boldsymbol{M}'\boldsymbol{1}_n\frac{1}{n}=\boldsymbol{\mu}.$$

设$\boldsymbol{Y}'\boldsymbol{a}$是$\boldsymbol{\mu}$的任一无偏估计,有

$\boldsymbol{E Y'a}=\boldsymbol{M'a}=\sum_1^n a_i\boldsymbol{\mu}=\boldsymbol{\mu}$，其充要条件为 $\Sigma a_i=1$。注意到

$$\begin{aligned}&\operatorname{Cov}\left(\boldsymbol{Y'a}-\boldsymbol{Y'1}\frac{1}{n},\ \boldsymbol{Y'1}\frac{1}{n}\right)\\&=\operatorname{Cov}\left(\sum_1^n\left(a_\alpha-\frac{1}{n}\right)\boldsymbol{y}_\alpha,\ \sum_1^n\frac{1}{n}\boldsymbol{y}_\alpha\right)\\&=\sum_1^n\frac{1}{n}\left(a_\alpha-\frac{1}{n}\right)\boldsymbol{\Sigma}=0.\end{aligned}$$

从而易见 $\operatorname{Cov}(\boldsymbol{Y'a})=\operatorname{Cov}(\boldsymbol{Y'a}-\overline{\boldsymbol{y}})+\operatorname{Cov}\overline{\boldsymbol{y}}$，得

$$\operatorname{Cov}(\boldsymbol{Y'a})-\operatorname{Cov}\overline{\boldsymbol{y}}=\operatorname{Cov}(\boldsymbol{Y'a}-\overline{\boldsymbol{y}})\geqslant 0.$$

顺便注意到 $\operatorname{Cov}\overline{\boldsymbol{y}}=\frac{1}{n}\boldsymbol{\Sigma}$.

由于

$$\begin{aligned}(n-1)E\boldsymbol{S}&=E\sum_{\alpha=1}^n(\boldsymbol{y}_\alpha-\overline{\boldsymbol{y}})(\boldsymbol{y}_\alpha-\overline{\boldsymbol{y}})'\\&=E\left(\sum_{\alpha=1}^n\boldsymbol{y}_\alpha\boldsymbol{y}'_\alpha-n\overline{\boldsymbol{y}}\overline{\boldsymbol{y}}'\right)\\&=\sum_{\alpha=1}^n(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}')-n\left(\frac{1}{n}\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}'\right)=(n-1)\boldsymbol{\Sigma},\end{aligned}$$

得 $\boldsymbol{S}$ 是 $\boldsymbol{\Sigma}$ 的无偏估计.

注记：定理 7.7 的证明并未涉及分布的具体形式，因而是不依赖于分布的.

定理 7.8

如果 $\boldsymbol{\Sigma}>0$，则 $(\overline{\boldsymbol{y}},\ \boldsymbol{S})$ 对于分布族 $\{N_{np}(\boldsymbol{\mu},\ \boldsymbol{I},\ \boldsymbol{\Sigma})\}$ 是充分完备统计量.

因此，$(\overline{\boldsymbol{y}},\boldsymbol{S})$ 是 $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的一致极小方差无偏估计(UMVUE).

证 Y 的联合密度是

$$\begin{aligned}f(\boldsymbol{Y})&=C\cdot(\det\boldsymbol{\Sigma})^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{\alpha=1}^n(\boldsymbol{y}_\alpha-\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_\alpha-\boldsymbol{\mu})\right\}\\&=C\cdot(\det\boldsymbol{\Sigma})^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\sum_{\alpha=1}^n(\boldsymbol{y}_\alpha-\boldsymbol{\mu})(\boldsymbol{y}_\alpha-\boldsymbol{\mu})'\right\}.\end{aligned}$$

由于

$$\sum_{\alpha=1}^{n}(\boldsymbol{y}_\alpha-\boldsymbol{\mu})(\boldsymbol{y}_\alpha-\boldsymbol{\mu})'=\sum_{\alpha=1}^{n}\boldsymbol{y}_\alpha\boldsymbol{y}_\alpha'-n\overline{\boldsymbol{y}}\boldsymbol{\mu}'-n\boldsymbol{\mu}\overline{\boldsymbol{y}}'+n\boldsymbol{\mu}\boldsymbol{\mu}'$$
$$=\boldsymbol{Y}'\boldsymbol{Y}-n\overline{\boldsymbol{y}}\boldsymbol{\mu}'-n\boldsymbol{\mu}\overline{\boldsymbol{y}}'+n\boldsymbol{\mu}\boldsymbol{\mu}'. \tag{7.16}$$

得 $$f(\boldsymbol{Y})=C\cdot(\det\boldsymbol{\Sigma})^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'\boldsymbol{Y}+n\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{y}}\right\}$$
$$\times\exp\left\{-\frac{n}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\}$$

可见$(\overline{\boldsymbol{y}},\ \boldsymbol{Y}'\boldsymbol{Y})$是充分完备统计量.(看定理 2.3.)

因为$(\overline{\boldsymbol{y}},\ \boldsymbol{S})$是$(\overline{\boldsymbol{y}},\ \boldsymbol{Y}'\boldsymbol{Y})$的函数,且是$(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$的无偏估计,故由定理 2.2,知其是 UMVUE.

定理 7.9

设 $\boldsymbol{\Sigma}>0$,则$(\overline{\boldsymbol{y}},\ \boldsymbol{W}/n)$是$(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$的极大似然估计. 这里 $\boldsymbol{W}=\boldsymbol{Y}'\boldsymbol{P}_{1^\perp}\boldsymbol{Y}$.

证 由(7.16),对于给定的 $\boldsymbol{\Sigma}$,欲使

$$f(\boldsymbol{Y};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma})$$

达极大,只需取 $\hat{\boldsymbol{\mu}}=\overline{\boldsymbol{y}}$,从而有

$$f(\boldsymbol{Y};\hat{\boldsymbol{\mu}},\ \boldsymbol{\Sigma})=C\cdot(\det\boldsymbol{\Sigma})^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\right\}.$$

故得

$$\log f(\boldsymbol{Y};\hat{\boldsymbol{\mu}},\ \boldsymbol{\Sigma})=\log C+\frac{n}{2}\log\det\boldsymbol{\Sigma}^{-1}-\frac{n}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}/n=\log C$$
$$-\frac{n}{2}\log\det\boldsymbol{W}/n+\frac{n}{2}(\log\det\boldsymbol{\Sigma}^{-1}\boldsymbol{W}/n$$
$$-\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}/n).$$

记 $\boldsymbol{\Sigma}^{-1}\boldsymbol{W}/n$ 的特征值为 $\lambda_1,\ \cdots,\ \lambda_p$. 则有

$$\log\det\boldsymbol{\Sigma}^{-1}\boldsymbol{W}/n-\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}/n$$
$$=\sum_1^p(\log\lambda_i-\lambda_i)=\sum_1^p[\log(1+\lambda_i-1)-\lambda_i]$$
$$\leqslant\sum_1^p(\lambda_i-1-\lambda_i)=-p$$

且上述不等式仅当 $\lambda_i=1,\ i=1,\ \cdots,\ p$ 时有等号成立. 因此 $f(\boldsymbol{Y};$

$\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}$)的极大值点是 $\hat{\boldsymbol{\Sigma}}=\boldsymbol{W}/n$. 证毕.

为了讨论 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的检验问题,先给出 $\bar{\boldsymbol{y}}$ 和 $\boldsymbol{W}$ 的分布.

定理 7.10

设 $\boldsymbol{Y}\sim N_{np}(\boldsymbol{M}, \boldsymbol{I}, \boldsymbol{\Sigma})$, $\boldsymbol{M}'=(\boldsymbol{\mu}\cdots\boldsymbol{\mu})$. $\bar{\boldsymbol{y}}=\boldsymbol{Y}'\boldsymbol{1}_n\dfrac{1}{n}$, $\boldsymbol{W}=\boldsymbol{Y}'\boldsymbol{P}_{1^\perp}\boldsymbol{Y}$. 则有

1° $\bar{\boldsymbol{y}}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$;

2° $\boldsymbol{W}\sim \boldsymbol{W}_p(n-1, \boldsymbol{\Sigma})$;

3° $\bar{\boldsymbol{y}}$ 与 $\boldsymbol{W}$ 独立.

证 1° 显然, 2° 由定理 7.3.1° 得. 令 $\boldsymbol{u}_n=\boldsymbol{1}_n\cdot\dfrac{1}{\sqrt{n}}$, 作正交阵 $\boldsymbol{U}=(\boldsymbol{u}_1\cdots\boldsymbol{u}_n)$, 考虑变换 $\boldsymbol{Y}=\boldsymbol{U}\boldsymbol{X}$, 则有 $\boldsymbol{W}=\boldsymbol{Y}'\boldsymbol{P}_{1^\perp}\boldsymbol{Y}=\boldsymbol{X}'\boldsymbol{U}'\boldsymbol{P}_{1^\perp}\boldsymbol{U}\boldsymbol{X}=\boldsymbol{X}'\begin{pmatrix}\boldsymbol{I}_{n-1} & \boldsymbol{0}\\ \boldsymbol{0} & 0\end{pmatrix}\boldsymbol{X}=\boldsymbol{X}_1'\boldsymbol{X}_1$, 其中 $\boldsymbol{X}_1$ 是 $\boldsymbol{X}$ 的前 $n-1$ 行. 而 $\bar{\boldsymbol{y}}=\boldsymbol{X}'\boldsymbol{U}'\boldsymbol{1}\dfrac{1}{n}=\dfrac{1}{\sqrt{n}}\boldsymbol{x}_n$ 这里 $\boldsymbol{x}_n'$ 是 $\boldsymbol{X}$ 的最后一行.

注意到 $\boldsymbol{X}'=\boldsymbol{Y}'\boldsymbol{U}$, 记 $\boldsymbol{X}'=(\boldsymbol{x}_1\cdots\boldsymbol{x}_n)$, 有

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)&=\mathrm{Cov}\left(\sum_{k=1}^n u_{k\alpha}\boldsymbol{y}_k, \sum_{k=1}^n u_{k\beta}\boldsymbol{y}_k\right)\\&=\sum_{k=1}^n u_{k\alpha}u_{k\beta}\mathrm{Cov}(\boldsymbol{y}_k)\\&=\delta_{\alpha\beta}\boldsymbol{\Sigma}, \quad \delta_{\alpha\beta}=\begin{cases}1, & \alpha=\beta;\\ 0, & \alpha\neq\beta.\end{cases}\end{aligned}$$

由于 $\boldsymbol{W}=\sum\limits_{\alpha=1}^{n-1}\boldsymbol{x}_\alpha\boldsymbol{x}_\alpha'$, $\bar{\boldsymbol{y}}=\dfrac{1}{\sqrt{n}}\boldsymbol{x}_n$, 3° 得证.

现在讨论检验假设

$$H_0: \mu=\mu_0.$$

先设 $\boldsymbol{\Sigma}>0$ 是已知的. 则可取检验统计量

$$T\triangleq n(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0).$$

显然有 $T\sim\chi_n^2(\delta)$, $\delta^2=n(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)$. 因此, 水平为 α 的拒绝域为

$$\{T\geqslant\chi_{n,\alpha}^2\}.$$

现设 $\boldsymbol{\Sigma}>0$ 未知. 我们用似然比去导出检验统计量. 令似然比

$$\lambda \triangleq \frac{M}{M_H}=\frac{\max\{L(\boldsymbol{Y};\boldsymbol{\mu},\boldsymbol{\Sigma}):\boldsymbol{\mu},\boldsymbol{\Sigma}\}}{\max\{L(\boldsymbol{Y};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}):\boldsymbol{\Sigma}\}},$$

由定理 7.8 的推导可见

$$\lambda=\frac{(\det n\boldsymbol{W}^{-1})^{\frac{n}{2}}\exp\left\{-\frac{n}{2}\right\}}{(\det n(\boldsymbol{W}+n(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)')^{-1})^{\frac{n}{2}}\exp\left\{-\frac{n}{2}\right\}}$$

$$=\left[\frac{\det(\boldsymbol{W}+n(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)')}{\det\boldsymbol{W}}\right]^{\frac{n}{2}}.$$

令

$$D=\begin{pmatrix}\boldsymbol{W} & \bar{\boldsymbol{y}}-\boldsymbol{\mu}_0\\ (\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)' & -\frac{1}{n}\end{pmatrix},$$

由附录 A.5.1 得

$$\det\boldsymbol{D}=\det\boldsymbol{W}\cdot\left(-\frac{1}{n}-(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)\right)$$

$$=-\frac{1}{n}\det(\boldsymbol{W}+n(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)').$$

从而有

$$\lambda=[1+n(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)]^{\frac{n}{2}}.$$

记

$$T^2=n(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0)'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_0) \tag{7.17}$$

知 λ 是 T^2 的严增函数, 故得似然比检验的拒绝域是 $\{T^2\geqslant C\}$. 由定理 7.6.2° 得: 当 H_0 成立, 有

$$T^2\cdot\frac{n-p}{p}\sim F_{p,n-p}.$$

因此取临界值 $C=\frac{p}{n-p}F_{p,n-p,\alpha}$, 就得水平为 α 的检验.

容易看出 T^2-检验正是一元统计中 t-检验的推广.

下面讨论 k 个总体情形的均值相等性检验.

设有 k 个独立正态总体, 分布为 $N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$, $i=1,\cdots,k$, $\boldsymbol{\Sigma}>0$ 未知. 欲检验假设

$$H_0:\boldsymbol{\mu}_1=\cdots=\boldsymbol{\mu}_k. \tag{7.18}$$

设自总体 $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ 抽取容量为 n_i 的样本

$$\boldsymbol{Y}^{(i)}=\begin{pmatrix}\boldsymbol{y}_1'^{(i)}\\ \vdots \\ \boldsymbol{y}_{n_i}'^{(i)}\end{pmatrix},\quad 记\ \boldsymbol{Y}=\begin{pmatrix}\boldsymbol{Y}^{(1)}\\ \vdots \\ \boldsymbol{Y}^{(k)}\end{pmatrix},\quad n=\sum_1^k n_i.$$

有似然函数

$$L(\boldsymbol{Y}, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$
$$=C\cdot(\det\boldsymbol{\Sigma})^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^k\sum_{\alpha=1}^{n_i}(\boldsymbol{y}_\alpha^{(i)}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_\alpha^{(i)}-\boldsymbol{\mu}_i)\right\}.$$

仿前算得似然比

$$\lambda=\frac{\max\{L: \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\}}{\max\{L: \boldsymbol{\mu}_1=\cdots=\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\}}=\left[\frac{\det(\boldsymbol{W}_1+\cdots+\boldsymbol{W}_k)}{\det\boldsymbol{W}}\right]^{-\frac{n}{2}}$$

其中 $\boldsymbol{W}_i=\boldsymbol{Y}^{(i)\prime}\boldsymbol{P}_{1_{n_i}^\perp}\boldsymbol{Y}^{(i)}$, $i=1, \cdots, k$, $\boldsymbol{W}=\boldsymbol{Y}'\boldsymbol{P}_{1_n^\perp}\boldsymbol{Y}$.

记
$$\Lambda=\frac{\det(\boldsymbol{W}_1+\cdots+\boldsymbol{W}_k)}{\det\boldsymbol{W}}.$$

注意到

$$\boldsymbol{W}=\sum_{i=1}^k\sum_{\alpha=1}^{n_i}(\boldsymbol{y}_\alpha^{(i)}-\overline{\boldsymbol{y}})(\boldsymbol{y}_\alpha^{(i)}-\overline{\boldsymbol{y}})'$$
$$=\sum_{i=1}^k\sum_{\alpha=1}^{n_i}(\boldsymbol{y}_\alpha^{(i)}-\overline{\boldsymbol{y}}^{(i)})(\boldsymbol{y}_\alpha^{(i)}-\overline{\boldsymbol{y}}^{(i)})'+\sum_{i=1}^k n_i(\overline{\boldsymbol{y}}^{(i)}-\overline{\boldsymbol{y}})(\overline{\boldsymbol{y}}^{(i)}-\overline{\boldsymbol{y}})'$$
$$\triangleq\sum_{i=1}^k\boldsymbol{W}_i+\boldsymbol{B},$$

显然有 $\sum_1^k\boldsymbol{W}_i$ 与 $\boldsymbol{B}$ 独立, $W\sim W_p(n-1, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$, $\boldsymbol{W}_i\sim W_p(n_i-1, \boldsymbol{\Sigma})$, $i=1, \cdots, k$. 由 Cochran 定理到 Wishart 分布情形的推广(看定理 7.5), 可得 $B\sim W_p(k-1, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$. 因此, 在假设 H_0 成立时, $\boldsymbol{\Delta}=0$, 有 $\Lambda\sim\Lambda(p, n-k, k-1)$, 得水平为 α 的检验的拒绝域 $\{\Lambda\leqslant\lambda_\alpha\}$中的 λ_α 可由查表得到.

现在研究 $k=2$ 时的特殊情形. 此时

$$B=\sum_1^2 n_i(\overline{\boldsymbol{y}}^{(i)}-\overline{\boldsymbol{y}})(\overline{\boldsymbol{y}}^{(i)}-\overline{\boldsymbol{y}})'$$
$$=\sum_1^2 n_i\left(\overline{\boldsymbol{y}}^{(i)}-\frac{n_i}{n}\overline{\boldsymbol{y}}^{(i)}-\frac{n_{3-i}}{n}\overline{\boldsymbol{y}}^{(3-i)}\right)$$

$$\times\left(\bar{\boldsymbol{y}}^{(i)}-\frac{n_i}{n}\bar{\boldsymbol{y}}^{(i)}-\frac{n_{3-i}}{n}\bar{\boldsymbol{y}}^{(3-i)}\right)'$$

$$=\sum_1^2\frac{n_i^2 n_{3-i}}{n^2}(\bar{\boldsymbol{y}}^{(i)}-\bar{\boldsymbol{y}}^{(3-i)})(\bar{\boldsymbol{y}}^{(i)}-\bar{\boldsymbol{y}}^{(3-i)})'$$

$$=\frac{n_1 n_2}{n}(\bar{\boldsymbol{y}}^{(1)}-\bar{\boldsymbol{y}}^{(2)})(\bar{\boldsymbol{y}}^{(1)}-\bar{\boldsymbol{y}}^{(2)})'.$$

仿(7.17)的推导，有 Λ 是 $T^2=\frac{n_1n_2}{n}(\bar{\boldsymbol{y}}^{(1)}-\bar{\boldsymbol{y}}^{(2)})'W^{-1}(\bar{\boldsymbol{y}}^{(1)}-\bar{\boldsymbol{y}}^{(2)})$ 的严降函数，得水平 α 的拒绝域$\left\{T^2\geqslant\frac{p}{n-p-1}F_{p,n-p-1,\alpha}\right\}$.

最后讨论正态总体$N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$的协方差阵 $\boldsymbol{\Sigma}$ 的检验问题．设 $\boldsymbol{Y}$ 为容量为 n 的样本．欲检验假设

$$H_0:\ \boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0>0(\text{已知}).$$

令似然比
$$\lambda=\frac{\max\{L(\boldsymbol{Y},\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}):\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}>0\}}{\max\{L(\boldsymbol{Y},\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}_0):\ \boldsymbol{\mu}\}}.$$

仿前可得

$$\lambda=\frac{\left(\det\frac{\boldsymbol{W}}{n}\right)^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}\right\}}{(\det\boldsymbol{\Sigma}_0)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{W}\right\}}$$

$$=\exp\left\{-\frac{n}{2}\right\}n^{\frac{np}{2}}(\det\boldsymbol{W}\boldsymbol{\Sigma}_0^{-1})^{-\frac{n}{2}}\exp\left\{\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{W}\right\}.$$

一般用 λ 的极限分布来作检验．由定理 3.5 知 $2\log\lambda\to\chi^2_{p(p+1)/2}$，当 $n\to\infty$．因此，当 n 较大时可近似地给出水平为 α 的拒绝域为

$$\{2\log\lambda\geqslant\chi^2_{p(p+1)/2,\alpha}\}.$$

从上述讨论可以看出，在多元统计分析中假设检验所遇到的一大困难是统计量的精确分布往往难以求出．从而，对检验的优良性就更加难以讨论．这方面的某些结果，如 T^2-检验的优良性，可参看 Giri «Multivariate Statistical Inference».

§7.2 判别分析

(一)判别的概念

分类问题是理论和实际工作者都会经常遇到的．当我们面临

一群对象的时候，只要想进行稍微深入一些的研究，就往往需要将对象分类，严格说来有两种类型的分类问题。一种是对这群对象的属类知之甚少，这时分类就带有较大程度的主观因素，例如可以依据对象的某些指标，将它们分类，也可以依据对象的另外一些指标，将它们分成另一些类．这种分类有相当的任意性．另一种分类是已知对象来自 k 个不同的类，我们的任务是将各个对象分别分到这 k 个类中的某一个去．这时分类的类别是已经确定了的．为了区别这两种分类，我们把前一种称为聚类，而把后一种称为判别．

随着多元分析的发展，聚类和判别已逐渐区分开来，并各自形成了一些实用的方法．但就理论而言，聚类分析比起其它分支来，还是很不完整的，限于本书的宗旨，就不准备在此介绍了．

判别问题的较确切的提法是：设有 k 个总体，记为 $G_1, \cdots, G_k$．已知样本来自这 k 个总体的某一个，但不知道它究竟来自其中哪一个．判别分析就是要根据对这 k 个总体的知识（由过去经验获得，或从这 k 个总体中抽样推断）和待判别的样本的一些指标的观察值，去作出样本来自某一总体的具体判断．

需要进行判别的问题，在现实生活中屡见不鲜．例如一个来医院就诊的病人，已发现有许多症状，需要确诊他患有何种疾病，如在地质考察中发现某地有一种矿石，需要判别是何种矿藏，并进而得出是贫矿还是富矿等结论；又如植物分类、社会调查、考古等各个性质很不相同的部门，都存在判别问题．因此，判别分析是一种很有实用意义的统计方法．

在判别分析中通常要考察样本的多个指标．例如，要判别患者的疾病究竟是不是肺炎，单纯用体温这项指标显然是不够的，还需要进行透视、化验、听诊等多项指标的考察，才有可能作出比较可靠的判别．因此，判别分析一般属于多元分析所讨论的范围．

进行判别的一个先决条件，是要对所涉及的 k 个总体有相当的了解，用统计的术语来说，就是要有关于这些总体的分布的知识．例如在考古学中要从头骨去确定它所属的人种，事先就要对各

种人种的头骨进行必要的测量，并得到充分数据，从而大致上获知总体的统计特性，以作为判别的依据.

下面我们介绍几种常用的判别方法. 其中有的方法较多地依赖统计直观，从理论角度看似不够严谨，但在总体分布的具体形式未知的条件下，就判别而言，难以有较理想的方法.

(二)距离判别

距离判别是一种很直观的判别方法，它可从不同角度导出.

例如我们可从假设检验的角度去讨论. 设 G_1, G_2 是两个不同的总体，它们分别有密度函数 f_1 和 f_2, 要判别 y 属于 G_1 或 G_2, 就相当于检验 y 的密度函数 f 是 f_1 或 f_2. 记

$$H_1: f=f_1,\ H_2: f=f_2,$$

于是判别问题就是去定出一个规则，根据观察值 y，究竟接受 H_1 还是 H_2. 按照似然比检验的直观想法，记 $\lambda=f_1(y)/f_2(y)$，当 $\lambda\leqslant C\ (C>0)$时，拒绝 H_1，否则就拒绝 H_2. 在假设检验中，这里的 C 是按照保护零假设的想法选择的，但在判别问题中一般不存在比较重视某一总体的理由，这时可认为假设 H_1, H_2 是地位对称的，从而选择 C 使得犯两类错误的概率相同. 如实际样本来自总体 G_j，但被误判属于 G_i，记这类误判概率为 p_{ij}. 此时要求 $p_{12}=p_{21}$.

判别规则可由判别域给定. 由上述方法导出的判别规则和判别域是

$$D_1=\left\{\boldsymbol{y}:\ \frac{f_1(\boldsymbol{y})}{f_2(\boldsymbol{y})}\leqslant C\right\},\ D_2=\left\{\boldsymbol{y}:\ \frac{f_1(\boldsymbol{y})}{f_2(\boldsymbol{y})}>C\right\},\quad (7.19)$$

当 $\boldsymbol{y}\in D_i$，判 $\boldsymbol{y}\in G_i$, $i=1, 2$.

下面给出这一方法在不同情形的应用.

1. 正态同协方差阵情形　设总体 G_1、G_2 的分布分别是 $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ 和 $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}>0$. 于是有

$$f_i(\boldsymbol{y})=C\cdot(\det\boldsymbol{\Sigma})^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_i)\right\},\ i=1, 2.$$

因此

$$\log\lambda=\log\frac{f_1(y)}{f_2(y)}=\frac{1}{2}[(\boldsymbol{y}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_2)-(\boldsymbol{y}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_1)]. \tag{7.20}$$

记 $W(\boldsymbol{y})=\log\lambda$, 称作**判别函数**, 它给出判别为

$$D_1=\{\boldsymbol{y}: W(\boldsymbol{y})\leqslant\log C\},\ D_2=\{\boldsymbol{y}: W(\boldsymbol{y})>\log C\}, \tag{7.21}$$

当 $\boldsymbol{y}\in D_i$, 判 $\boldsymbol{y}\in G_i$, $i=1, 2$.

经简单计算有

$$W(\boldsymbol{y})=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\overline{\boldsymbol{\mu}}),\ \text{其中}\ \overline{\boldsymbol{\mu}}=\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2).$$

在两总体地位对称时, 我们取 $C=1$(理由可由下面的讨论自明). 这时**判别规则**可由判别函数描述为

$$\begin{cases} W(\boldsymbol{y})\leqslant 0, \text{判}\ \boldsymbol{y}\ \text{属}\ G_1\ \text{总体}, \\ W(\boldsymbol{y})>0, \text{判}\ \boldsymbol{y}\ \text{属}\ G_2\ \text{总体}. \end{cases} \tag{7.22}$$

(注意, 当两个总体都有连续型分布, $\boldsymbol{W}(\boldsymbol{y})=0$ 这一事件出现的概率为 0, 这种情形可予忽略, 原则上此时可判 y 属任一总体, 故判它属 G_2 也无妨.)

从(7.20)不难看出, 判别函数(不计常数因子)是二次型的差, 而这种类型的二次型

$$(\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y}) \tag{7.23}$$

和欧氏空间 R^n 中距离概念的一个重要推广相联系, 那就是 1936 年由 Mahalanobis 所引进的**马氏距离**

$$d(\boldsymbol{x},\ \boldsymbol{y})\triangleq[(\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})]^{\frac{1}{2}}. \tag{7.24}$$

亦记 $d(\boldsymbol{x},\ \boldsymbol{y})\triangleq\|\boldsymbol{x}-\boldsymbol{y}\|_{\Sigma}$, 称 $\boldsymbol{\Sigma}$ 为**权矩阵**.

定义 $\boldsymbol{y}$ 到总体 G_i 的距离为

$$d(\boldsymbol{y},\ G_i)=\|\boldsymbol{y}-\boldsymbol{\mu}_i\|_{\Sigma},\ i=1,\ 2, \tag{7.25}$$

从而, 判别规则(7.22)可改述为

$$\begin{cases} \text{当}\ d(\boldsymbol{y},\ G_1)<d(\boldsymbol{y},\ G_2), \text{判}\ \boldsymbol{y}\in G_1, \\ \text{当}\ d(\boldsymbol{y},\ G_1)\geqslant d(\boldsymbol{y},\ G_2), \text{判}\ \boldsymbol{y}\in G_2. \end{cases} \tag{7.26}$$

所以, 我们称(7.22)或(7.26)给出的判别为**距离判别**, 即若 $\boldsymbol{y}$ 离

总体 G_i 较近. 就判 $\boldsymbol{y}$ 属 G_i.

距离判别中集合

$$\{y: d(y, G_1)=d(y, G_2)\} \tag{7.27}$$

称为**判别的边界**. 在两总体有相同协方差阵且遵从正态分布情形, 此边界为过 $\bar{\mu}$ 的超平面, 称 $\bar{\mu}$ 是判别的阈值.

下面对 $C>0$ 讨论误判概率. 记 $\boldsymbol{b}=\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$, (由7.21), 当 $\boldsymbol{y}$ 来自总体 G_1 有

$$W(\boldsymbol{y})\sim N_p\left(\frac{1}{2}\|\boldsymbol{b}\|^2, \|\boldsymbol{b}\|^2\right),$$

得

$$\begin{aligned}p_{21}&=p_1(W(\boldsymbol{y})\leqslant\log C)\\&=p_1\left(\frac{W(\boldsymbol{y})-E(W(\boldsymbol{y}))}{\sqrt{D(W(\boldsymbol{y}))}}\leqslant\frac{\log C-\frac{1}{2}\|\boldsymbol{b}\|^2}{\|\boldsymbol{b}\|}\right)\\&=\phi\left(\left(\log C-\frac{1}{2}\|\boldsymbol{b}\|^2\right)\Big/\|\boldsymbol{b}\|\right),\end{aligned} \tag{7.28}$$

其中 ϕ 是 $N(0, 1)$ 的分布函数.

相仿可得, 当 $\boldsymbol{y}$ 来自总体 G_2 而误判入总体 G_1 的概率是

$$p_{12}=1-\phi\left(\left(\log C+\frac{1}{2}\|\boldsymbol{b}\|^2\right)\Big/\|\boldsymbol{b}\|\right). \tag{7.29}$$

注意到 p_{21}, p_{12} 都是 $\|\boldsymbol{b}\|$ 的严降函数, 而 $\|\boldsymbol{b}\|^2$ 为

$$\|\boldsymbol{b}\|^2=\boldsymbol{b}'\boldsymbol{b}=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\triangleq d^2(G_1, G_2)$$

可视为**两总体的马氏距离**. 因此, 当两总体距离较远, 则误判概率都较小, 否则就较大, 当两总体很接近时, 有 $p_{12}+p_{21}\approx 1$. 并且有

$$\begin{cases}C=1, & p_{21}=p_{12};\\C>1, & p_{21}>p_{12};\\C<0, & p_{21}<p_{12}.\end{cases} \tag{7.30}$$

故可由 C 的选择来控制 p_{21} 与 p_{12} 的差异.

为了使(7.26)确定的判别规则实际可行, 就必须知道两总体的均值和协方差阵, 而这往往是不大可能做到的, 故在实际中不得不先对 $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ 和 $\boldsymbol{\Sigma}$ 作出估计. 一般地, 可取 G_i 中容量为 n_i 的样

本 $\boldsymbol{Y}^{(i)}$, $i=1, 2$. 由 § 7.1.(三), 以估计

$$\hat{\boldsymbol{\mu}}_i=\frac{1}{n_i}\boldsymbol{Y}'^{(i)}\mathbf{1}_{n_i},\ i=1,\ 2,\ \hat{\boldsymbol{\Sigma}}=\frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\boldsymbol{Y}'^{(i)}\boldsymbol{P}_{\mathbf{1}_{n_i}^{\perp}}\boldsymbol{Y}^{(i)}$$

代 $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}$ 得

$$\hat{W}(\boldsymbol{y})=(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)'(\hat{\boldsymbol{\Sigma}})^{-1}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2/2). \tag{7.31}$$

以 $\hat{W}(\boldsymbol{y})$代替判别规则(7.22)中的 $W(\boldsymbol{y})$, 可得(基于样本观察值的)距离判别. 对于它的误判概率的研究就变得相当复杂,不在此进行.

2. 分布自由情形　现在不限制总体的分布形式, 只假定 G_i 有相同协方差阵 $\boldsymbol{\Sigma}$, 有均值 $\boldsymbol{\mu}_i$, $i=1, 2$, $\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$.

注意到(7.22)给出的判别规则,本身只和分布的二阶矩有关,依然可看作是由样本到总体的马氏距离的大小来作出判别. 我们可将此距离判别移用到分布自由情形中来. 此时, 因不知道分布形式,当然无法计算误判概率的大小,故所得规则可看成一种实用的方法,方法的优良性也唯有用实践去检验.

容易想到一个问题: 既然是用距离判别,为什么一定要用马氏距离呢? 其它距离可不可用? 欧氏距离岂不更简单? 要确切地回答这些问题是困难的. 事实上,采用其它距离的方法已在实际工作中出现. 下面仅就马氏距离和欧氏距离的比较,谈一些看法. 首先,在实际问题中必然涉及尺度的选择,各个指标一般有不同的属性和量纲. 如果采用欧氏距离,不难看出,它和量纲的选取有很大关系, 那么, 结果将随量纲的不同而可能很不相同, 这明显是不合理的. 但马氏距离是无量纲的,它受尺度选取的影响就自然较少,这是用马氏距离的理由之一. 另外, 欧氏距离未把各指标的方差和相关考虑进去, 而马氏距离中的 Σ 是协方差阵, 从而把方差和相关考虑在内, 结果可以使两者相去甚远. 例如 $\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}$, $\sigma_1^2>\sigma_2^2$, 欧氏距离为

$$\|\boldsymbol{x}-\boldsymbol{y}\|=[(x_1-y_1)^2+(x_2-y_2)^2]^{\frac{1}{2}},$$

而马氏距离是

$$\|x-y\|_\Sigma=[(x_1-y_1)^2\sigma_1^{-2}+(x_2-y_2)^2\sigma_2^{-2}]^{\frac{1}{2}}.$$

因 $\sigma_1>\sigma_2$，第一个指标的方差大，其差异的意义就相对较小，故用 σ_1^2 除后再作为距离中的求和项，明显较为合理．又若 $\boldsymbol{\Sigma}=\begin{pmatrix}\sqrt{2} & 1\\ 1 & \sqrt{2}\end{pmatrix}$，则 $\boldsymbol{\Sigma}^{-1}=\begin{pmatrix}\sqrt{2} & -1\\ -1 & \sqrt{2}\end{pmatrix}$．令 $\boldsymbol{x}=(1,\ 1)'$，$\boldsymbol{y}=(1,\ -1)'$，如考虑它们到 0 的欧氏距离，有 $\|\boldsymbol{x}\|=\|\boldsymbol{y}\|=\sqrt{2}$．但马氏距离 $\|\boldsymbol{x}\|_\Sigma=\sqrt{2(\sqrt{2}-1)}<\sqrt{2(\sqrt{2}+1)}=\|\boldsymbol{y}\|_\Sigma$．由于两个指标存在正相关，故当两指标异号时与原点距离较远是合乎情理的．这就是支持马氏距离作为判别依据的又一理由．

3．*协方差矩阵不等的情形*　先设 G_1，G_2 是两个总体，分别有均值 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$ 和非奇异协方差阵 $\Sigma_1\neq\Sigma_2$．仿(7.20)令判别函数

$$W(\boldsymbol{y})=\frac{1}{2}[(\boldsymbol{y}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_2)-(\boldsymbol{y}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_1)] \tag{7.32}$$

仍然用判别规则(7.22)．但此时 $W(\boldsymbol{y})$ 不再是线性函数，判别边界将是 p 维空间中的二次曲面，误判概率的计算更为复杂．现仅以 $p=1$、总体分布为正态为例进行讨论：不妨设 $\mu_1>\mu_2$，$y\in(\mu_2,\ \mu_1)$，由 $d(y,\ G_1)=d(y,\ G_2)$ 可推出 $\frac{y-\mu_2}{\sigma_2}=\frac{\mu_1-y}{\sigma_1}$，解得

$$y_0=\frac{\mu_1\sigma_2+\mu_2\sigma_1}{\sigma_1+\sigma_2}.\quad(\neq\bar{\mu})$$

y_0 正是判别的阈值．判别可记为

$y\leqslant y_0$，判 $y\in G_2$；$y>y_0$，判 $y\in G_1$．

若 $\sigma_2>\sigma_1$，阈值较靠近 μ_1；若 $\sigma_2<\sigma_1$，阈值较靠近 μ_2．不难算出误判概率

$$p_{12}=\int_{-\infty}^{a}(2\pi\sigma_1^2)^{-\frac{1}{2}}\exp\left\{-\frac{(y-\mu_1)^2}{2\sigma_1^2}\right\}dy=\int_{-\infty}^{b}(2\pi)^{-\frac{1}{2}}\exp\left\{-\frac{x^2}{2}\right\}dx$$
$$=\phi\left(\frac{\mu_2-\mu_1}{\sigma_1+\sigma_2}\right),$$

这里，$a=\frac{\mu_1\sigma_2+\mu_2\sigma_1}{\sigma_1+\sigma_2}$，$b=\frac{\mu_2-\mu_1}{\sigma_1+\sigma_2}$

其中 ϕ 是 $N(0, 1)$的分布函数,相仿可得

$$p_{21}=1-\phi\left(\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2}\right)=p_{12}.$$

现在考虑 k 个总体的一般情形. 设 G_i 总体有均值 $\boldsymbol{\mu}_i$, 协方差阵 $\boldsymbol{\Sigma}_i$,记 $d(\boldsymbol{y}, G_i)=\|\boldsymbol{y}-\boldsymbol{\mu}_i\|_{\Sigma_i}$,且令

$$D_i=\{\boldsymbol{y}: d(\boldsymbol{y}, G_i)=\min_{1<j<k} d(\boldsymbol{y}, G_j)\}, \ i=1, \cdots, k. \tag{7.33}$$

为判别区域. 判别规则是

$$\text{当 } \boldsymbol{y}\in D_i, \text{判 } \boldsymbol{y} \text{ 来自 } G_i, \ i=1, \cdots, k. \tag{7.34}$$

此规则在 D_i 与 D_j 的交界处有不确定性,但若总体存在密度函数,在交界处可判属这两个总体的任一个. 因交集的概率为零,这种不确定性不妨予以忽略. D_i 的计算涉及 p 维空间中的二次曲面,原则上并不十分困难. 但我们不必求出它的界限,而可由 D_i 的定义,根据 k 个值$\{d(y, G_j), j=1, \cdots, k\}$的极小值就作出判别. 误判概率的计算不在此讨论.

(三)Fisher 判别函数

1. *最优线性判别函数* Fisher 提出了一个很有意思的想法,假定仅仅考虑线性判别函数 $\boldsymbol{a}'\boldsymbol{y}$, 即以各个指标的加权和作为一个综合指标来作为判别的依据. 那么,怎样的线性判别函数是最好的呢? Fisher 给出了如下准则:一个判别函数是好的,它在各个总体中去求均值所得 k 个数应有较大离差, 但这个离差的意义应相对于判别函数总方差去衡量,为此可令

$$\Delta(\boldsymbol{a})\triangleq\frac{\sum_1^k\left[E_i\boldsymbol{a}'\boldsymbol{y}-\frac{1}{k}\sum_1^k E_i\boldsymbol{a}'\boldsymbol{y}\right]^2}{\sum_1^k D_i(\boldsymbol{a}'\boldsymbol{y})} \tag{7.35}$$

其中 E_i, D_i 表示在第 i 个总体中求均值和方差. $\Delta(\boldsymbol{a})$被称为 **Fisher** 准则. 我们有

定义 7.2

如果 $\boldsymbol{u}'\boldsymbol{y}$ 满足

$$\Delta(\boldsymbol{u})=\max_{\boldsymbol{a}\in R^p}\Delta(\boldsymbol{a}) \tag{7.36}$$

则称 $\boldsymbol{u}'\boldsymbol{y}$ 是最优线性判别函数.

由矩阵代数知识，不难求出(7.36)的解. 将(7.35)表成矩阵形式有

$$\Delta(\boldsymbol{a})=\frac{\boldsymbol{a}'\sum_{1}^{k}(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})'\boldsymbol{a}}{\boldsymbol{a}'\left(\sum_{1}^{k}\boldsymbol{\Sigma}_i\right)\boldsymbol{a}}\triangleq\frac{\boldsymbol{a}'\boldsymbol{M}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}.$$

(很明显 $\Delta(\boldsymbol{a})$ 只依赖于 $\boldsymbol{a}$ 的方向，而与 $\boldsymbol{a}$ 的长度 $\|\boldsymbol{a}\|$ 无关.)其中 $\boldsymbol{\mu}_i$ 是第 i 个总体的均值，$\boldsymbol{\Sigma}_i$ 是第 i 个总体的协方差阵，$\overline{\boldsymbol{\mu}}=\frac{1}{k}\sum_{1}^{k}\boldsymbol{\mu}_i$.

$$\boldsymbol{M}=\sum_{1}^{k}(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i-\overline{\boldsymbol{\mu}})',\quad \boldsymbol{\Sigma}=\sum_{1}^{k}\boldsymbol{\Sigma}_i,$$ 设 $\boldsymbol{\Sigma}_i>0$. 令 $\boldsymbol{b}=\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{a}$, 不妨设 $\|\boldsymbol{b}\|=1$, 于是有

$$\Delta(\boldsymbol{a})=\Delta(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b})=\boldsymbol{b}'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b}.$$

由附录 A.4.2 知,

$$\max_{\|\boldsymbol{b}\|=1}\boldsymbol{b}'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{b}=\lambda_1(\boldsymbol{\Sigma}^{-1}\boldsymbol{M})\triangleq\lambda_1$$

极大值点为 $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ 的相应于 λ_1 的特征向量 $\boldsymbol{c}_1$, 因此 $\boldsymbol{a}_1=\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{c}_1$ 是 $\boldsymbol{\Sigma}^{-1}\boldsymbol{M}$ 的相应于 λ_1 的特征向量，得最优线性判别函数 $W_1(\boldsymbol{y})=\boldsymbol{a}_1'\boldsymbol{y}$, 称 λ_1 是 $W_1(\boldsymbol{y})$ 的判别效率.

依据 $W_1(\boldsymbol{y})$ 的判别规则是: 计算 $W_1(\boldsymbol{y})$ 与 $W_1(\boldsymbol{\mu}_i)$ 的距离

$$d(W_1(\boldsymbol{y}),\ W_1(\boldsymbol{\mu}_i))=\frac{|\boldsymbol{a}_1'\boldsymbol{y}-\boldsymbol{a}_1'\boldsymbol{\mu}_i|}{(\boldsymbol{a}_1'\boldsymbol{\Sigma}_i\boldsymbol{a}_1)^{\frac{1}{2}}},\quad i=1,\ \cdots,\ k, \tag{7.37}$$

然后将 $\boldsymbol{y}$ 判属使(7.37)达最小值的那个总体.

由于 $\boldsymbol{\mu}_i$ 不全相同，$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ 至少有一个正特征值. 一般地说，$\boldsymbol{\Sigma}^{-1}\boldsymbol{M}$ 的正特征值不至一个，设为 $\lambda_1\geqslant\cdots\geqslant\lambda_r>0$, 而用一个综合指标 $\boldsymbol{a}_1'\boldsymbol{y}$ 对 $\boldsymbol{y}$ 进行判别也可能不足以得出确切结论，这时可引进相继的最优线性判别函数，$W_i(\boldsymbol{y})=\boldsymbol{a}_i'\boldsymbol{y}$, $i=2,\ \cdots,\ r$. 其中 $\boldsymbol{a}_i$ 是 $\boldsymbol{\Sigma}^{-1}\boldsymbol{M}$ 的相应于特征值 λ_i 的特征向量，$i=2,\ \cdots,\ r$. 可用 r

个综合指标来判别 $\boldsymbol{y}$. 记 $A=(\boldsymbol{a}_1\cdots\boldsymbol{a}_r)$, 有 $\boldsymbol{A}'\boldsymbol{y}$ 到 G_i 的马氏距离为

$$d^2(\boldsymbol{A}'\boldsymbol{y},\ \boldsymbol{A}'\boldsymbol{\mu}_i)=(\boldsymbol{y}-\boldsymbol{\mu}_i)'\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{\Sigma}_i\boldsymbol{A})^{-1}\boldsymbol{A}'(\boldsymbol{y}-\boldsymbol{\mu}_i) \tag{7.38}$$

从而仿(7.34)可得距离判别的判别规则.

这种用 r 个指标 $\boldsymbol{A}'\boldsymbol{y}$ 代替 p 个指标 $\boldsymbol{y}$ 进行推断的思想, 与下节将要叙述的主成分分析的基本精神是一致的.

2. Fisher 准则与其它判别的关系　Fisher 准则与距离判别的关系, 在 $k=2$ 时易于看出. 此时

$$\Delta(\boldsymbol{a})=\frac{1}{2}\boldsymbol{a}'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{a}/2\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a},\ \boldsymbol{\Sigma}=\frac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2).$$

因此 $\boldsymbol{\Sigma}^{-1}\boldsymbol{M}=\frac{1}{2}\ \Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'$, 它有唯一特征值为 $\frac{1}{2}\times(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\frac{1}{2}\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2_{\Sigma}$, 相应的特征向量(不计常数因子)为

$\boldsymbol{a}_1=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$, 得最优线性判别函数

$$W_1(\boldsymbol{y})=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}. \tag{7.39}$$

故(7.37)给出的判别与(7.26)一致(在 $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$ 时).

不难将 Fisher 准则推广, 即在(7.35)中将 $\boldsymbol{a}'\boldsymbol{y}$ 换为 $\boldsymbol{y}$ 的任意(可测)函数 $a(\boldsymbol{y})$, 情形将会怎样呢? 设 $k=2$, G_i 有密度函数 $f_i(\boldsymbol{y})$, 于是 $E_i a=\int a(\boldsymbol{y})f_i(\boldsymbol{y})d\boldsymbol{y}$, 得

$$\Delta(a)=\frac{\left\{\int a(\boldsymbol{y})[f_1(\boldsymbol{y})-f_2(\boldsymbol{y})]d\boldsymbol{y}\right\}^2}{\int[a(\boldsymbol{y})-E_1\boldsymbol{a}]^2f_1(\boldsymbol{y})d\boldsymbol{y}+\int[a(\boldsymbol{y})-E_2\boldsymbol{a}]^2f_2(\boldsymbol{y})d\boldsymbol{y}}, \tag{7.40}$$

设 $\Delta(\boldsymbol{a})$ 的极大值点为 $a_0(\boldsymbol{y})$, 令 $a(\boldsymbol{y})=a_0(\boldsymbol{y})+\lambda b(\boldsymbol{y})$, 其中 $\lambda\in\boldsymbol{R}'$, $b(\boldsymbol{y})$ 是任意(可测)函数, 从而可记

$$\Delta(a)=\Delta(a_0,\ \lambda,\ b).$$

视 $\Delta(a)$ 为 λ 的函数, 应在 $\lambda=0$ 达极大值, 故有 $\left.\frac{\partial\Delta(a)}{\partial\lambda}\right|_{\lambda=0}=0$. 记

$$E_{i0}=\int a_0(\boldsymbol{y})f_i(\boldsymbol{y})d\boldsymbol{y},\ D_{i0}=\int [a_0(\boldsymbol{y})-E_{i0}]^2 f_i(\boldsymbol{y})d\boldsymbol{y},\ i=1,\ 2,$$

经初等计算得 $a_0(\boldsymbol{y})$ 应满足

$$\int\{(D_{10}+D_{20})(f_1-f_2)-(E_{10}-E_{20})[(a_0-E_{10})f_1 + (a_0-E_{20})f_2]\}b(\boldsymbol{y})d\boldsymbol{y}=0,$$

由 $b(\boldsymbol{y})$ 之任意性可得

$$(D_{10}+D_{20})(f_1-f_2)-(E_{10}-E_{20})[(a_0-E_{10})f_1 + (a_0-E_{20})f_2]=0,$$

解出

$$a_0(\boldsymbol{y})=\frac{(d+E_{10})f_1-(d-E_{20})f_2}{f_1+f_2},\quad 其中\ d=\frac{D_{10}+D_{20}}{E_{10}-E_{20}}. \tag{7.41}$$

由 $\Delta(a)$ 的表达式(7.40)易见，当 $a_0(\boldsymbol{y})$ 是极大值点，它的线性函数 $\alpha a_0(\boldsymbol{y})+\beta$ 亦是(7.40)的极大值点.

设 π_1, π_2 分别是总体 G_1 和 G_2 的先验概率，即在混合总体中，G_i 所占的比重为 $\pi_i(\pi_1+\pi_2=1)$, $i=1, 2$. 用待定系数法，容易算出，当取

$$\alpha=\frac{\pi_1+\pi_2}{2d+E_{10}-E_{20}},\ \beta=\frac{\pi_2(d-E_{20})-\pi_1(d+E_{10})}{2d+E_{10}-E_{20}},$$

就有

$$\alpha a_0(\boldsymbol{y})+\beta=\frac{\pi_2 f_1(\boldsymbol{y})\pi_1 f_2(\boldsymbol{y})}{f_1(\boldsymbol{y})+f_2(\boldsymbol{y})}\triangleq a(\boldsymbol{y}) \tag{7.42}$$

是(7.40)的极大值点，即它是最优判别函数. 以 $a(\boldsymbol{y})$ 为判别函数，得判别域

$$\begin{aligned} D_1&=\{\boldsymbol{y}: a(\boldsymbol{y})>0\}=\left\{\boldsymbol{y}: f_1(\boldsymbol{y})>\frac{\pi_1}{\pi_2}f_2(\boldsymbol{y})\right\}, \\ D_2&=\{\boldsymbol{y}: a(\boldsymbol{y})\leqslant 0\}=\left\{\boldsymbol{y}: f_1(\boldsymbol{y})\leqslant\frac{\pi_1}{\pi_2}f_2(\boldsymbol{y})\right\}. \end{aligned} \tag{7.43}$$

这就与(一)中由假设检验方法所得的判别(7.19)相同$\left(取\ C=\frac{\pi_1}{\pi_2}\ 时\right)$.

从下段的讨论还可看出，当我们用 Bayes 方法去研究判别时，

在总体个数 $k=2$ 时，所得结果亦与(7.43)一致。

(四) **Bayes 判别**

Bayes 判别是判别分析中最常用的方法之一。

Bayes 统计的基本思想已在第五章中阐述。将它用于判别无任何原则困难。在判别问题中，参数 θ 实际上只取 k 个值，分别对应着 k 个总体的密度 f_i, $i=1, \cdots, k$. 不妨认为参数 θ 就取值 $1, \cdots, k$. 设 $\pi_j \triangleq P(\theta=j)$, $j=1, \cdots, k$, 是先验概率，如果样本来自 G_j 但误判属 G_i 的损失记为 $L(i, j)$. 而这种情况发生的概率(即误判概率)为

$$p_{ij}=\int_{D_i} f_j(\boldsymbol{y})d\boldsymbol{y} \triangleq \int f_j(\boldsymbol{y})\chi_{D_i}(\boldsymbol{y})d\boldsymbol{y},\ D_i \text{ 是判别域.}$$

于是平均损失(即风险函数)是

$$\sum_{i=1}^{k} L(i, j)\int \chi_{D_i}(\boldsymbol{y})f_j(\boldsymbol{y})d\boldsymbol{y}.$$

这时判决函数实质上就相当于一组判别域

$$D=\{D_1, \cdots, D_k\}, \text{满足 } \cup_1^k D_i=R^p,\ P_h(D_i\cap D_j)=0, \tag{7.44}$$

其中 $i\neq j$, $i, j, h=1, \cdots, k$, 当 $y\in D_i$, 就判 y 来自 G_i.

由先验分布 $\boldsymbol{\pi}$ 得 Bayes 风险为

$$\begin{aligned} R_\pi(D) &= \sum_{j=1}^{k}\pi_j\sum_{i=1}^{k} L(i, j)\int\chi_{D_i}(\boldsymbol{y})f_j(\boldsymbol{y})d\boldsymbol{y} \\ &= \sum_{i=1}^{k}\int\chi_{D_i}(\boldsymbol{y})\sum_{j=1}^{k}\pi_j L(i, j)f_j(\boldsymbol{y})d\boldsymbol{y} \end{aligned} \tag{7.45}$$

记 $h_i(\boldsymbol{y})=\sum_{j=1}^{k}\pi_j L(i, j)f_j(y)$. 现在的目标就是求一组判别域 D 使得 $R_\pi(D)$ 达极小值，称 D 为 Bayes 解，或 **Bayes 判别**.

$$\text{令 } D_i=\{\boldsymbol{y}: h_i(\boldsymbol{y})=\min_{1\leq j\leq k} h_j(\boldsymbol{y})\},\ i=1, \cdots, k. \tag{7.46}$$

我们来证明这样给出的 D 就是 Bayes 解. 事实上，设 D^* 是任一组判别域，我们有

$$
\begin{aligned}
&R_\pi(D^*)-R_\pi(D)\\
&=\sum_{i=1}^{k}\int \chi_{D_i^*}(\boldsymbol{y})h_i(\boldsymbol{y})d\boldsymbol{y}-\sum_{i=1}^{k}\int \chi_{D_i}(\boldsymbol{y})h_i(\boldsymbol{y})d\boldsymbol{y}\\
&=\sum_{j=1}^{k}\sum_{i=1}^{k}\int \chi_{D_j}(\boldsymbol{y})\chi_{D_i^*}(\boldsymbol{y})h_i(\boldsymbol{y})d\boldsymbol{y}\\
&\quad-\sum_{j=1}^{k}\sum_{i=1}^{k}\int \chi_{D_j^*}(\boldsymbol{y})\chi_{D_i^*}(\boldsymbol{y})h_i(\boldsymbol{y})d\boldsymbol{y}\\
&=\sum_{j=1}^{k}\sum_{i=1}^{k}\int \chi_{D_j^*\cap D_i}(\boldsymbol{y})[h_j(\boldsymbol{y})-h_i(\boldsymbol{y})]d\boldsymbol{y},
\end{aligned}
$$

由于 D_i 的定义知 $\chi_{D_j^*\cap D_i}(\boldsymbol{y})[h_j(\boldsymbol{y})-h_i(\boldsymbol{y})]\geqslant 0$，故得 $R_\pi(D^*)-R_\pi(D)\geqslant 0$，因此 D 是 Bayes 解.

现设

$$
L(i,\ j)=\begin{cases}1,\ i\neq j,\\ 0,\ i=j.\end{cases}
$$

则有 Bayes 判别的判别域为

$$
D_i=\{\boldsymbol{y}:\ \sum_{j\neq i}\pi_j f_j(\boldsymbol{y})=\min_{1<l<k}\sum_{j\neq l}\pi_j f_j(\boldsymbol{y})\},\ i=1,\ \cdots,\ k. \tag{7.47}
$$

在 $k=2$ 的特殊情形下，由(7.47)得

$$
D_1=\left\{y:\ f_1(y)\geqslant\frac{\pi_1}{\pi_2}f_2(y)\right\},
$$

$$
D_2=\left\{y:\ f_1(y)\leqslant\frac{\pi_1}{\pi_2}f_2(y)\right\}
$$

与(7.43)一致. 记 $C=\pi_1/\pi_2$，又与判别(7.19)相同. 几种判别方法在一定条件下的一致，实际上就从不同角度说明了这些判别方法的合理性.

§7.3 多元线性模型

在一元统计分析中，我们在线性模型的范围内讨论了应用最广的回归分析和方差分析方法. 一元线性模型刻划了一个因变量对一个或一组自变量的统计依赖关系. 当这种统计依赖关系对于自变量的效应而言是线性的，或近似于线性的，用线性模型来处理

是方便而有效的. 在本节中，我们把线性模型自然地推广到多个因变量情形，并得到回归分析和方差分析的推广.

(一)多元线性模型

一般地说，在实际应用中，我们仅仅对试验结果中的单个指标感兴趣的情形是较少见的，更多的是关心 p 个相互联系着的指标，也就是 p 个随机变量 $Y_1, \cdots, Y_p$，称为因变量. 这些试验结果的观察值有可能线性地依赖作为试验条件的自变量 $X_1, \cdots, X_q$ 的效应，但存在着试验误差. 于是 $Y_1, \cdots, Y_p$ 就可表为这些效应的线性模型. 在不可控制的观测中，可化为线性模型的情形也很常见. 例如在某个地域的 n 个点上，测定了地质样品中 p 种化学元素式化合物的含量，得到 n 个 p 维向量 $y_1, \cdots, y_n$. 这些 y_α 应当是地理位置 (u, v) 的连续函数，可以用多项式去近似，如采用二次多项式，则 y_α 看作多项式系数的函数是线性的，考虑到观测与模型误差，我们有多元线性模型

$$\begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} \vdots & u_1 & v_1 & u_1^2 & v_1^2 & u_1v_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & u_n & v_n & u_n^2 & v_n^2 & u_nv_n \end{pmatrix} \begin{pmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \vdots & \vdots & & \vdots \\ \beta_{s1} & \beta_{s2} & \cdots & \beta_{sp} \end{pmatrix} + \begin{pmatrix} \varepsilon_1' \\ \vdots \\ \varepsilon_n' \end{pmatrix}. \tag{7.48}$$

这就是地质中趋势面分析的模型. 当然，为了对这些模型进行分析，还得附加一些必要和合理的假定.

现在我们给出一般的多元线性模型，从形式上讲，它只不过是将一元情形的一个因变量 Y 推广到多个因变量 $Y_1, \cdots, Y_p$. 试验点的概念和记法与一元情形完全相同，n 个试验点记为

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nq} \end{pmatrix},$$

也称 $\boldsymbol{X}$ 为设计矩阵. 但这里 n 个观察值 $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ 都是 p 维向量，

应记成 $n\times p$ 阶矩阵

$$\boldsymbol{Y}=\begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix}, \quad \text{相应的误差阵 } \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1p} \\ \vdots & & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{np} \end{pmatrix}.$$

未知参数矩阵记为

$$\boldsymbol{B}=\begin{pmatrix} \beta_{11} & \cdots & \beta_{1p} \\ \vdots & & \vdots \\ \beta_{q1} & \cdots & \beta_{qp} \end{pmatrix}, \text{是一个 } q\times p \text{ 阶阵.}$$

它的第 j 列 $\boldsymbol{\beta}_j$ 影响着试验结果的第 j 个指标. 若记 $\boldsymbol{Y}$ 的第 j 列为 $\boldsymbol{Y}_j$, 我们有模型

$$\boldsymbol{Y}_j=\boldsymbol{X\beta}_j+\boldsymbol{\varepsilon}_j,\ j=1,\ \cdots,\ p, \tag{7.49}$$

由(7.49), 多元线性模型似乎单纯是 p 个一元线性模型的混合. 这就涉及到我们一开始就提到过的 p 个指标之间理应存在相关, 设观察值矩阵的各行互不相关且有相同的协方差矩阵 Σ, 也就是说

$$\mathrm{Cov}(\boldsymbol{y}_\alpha,\ \boldsymbol{y}_\beta)=\delta_{\alpha\beta}\boldsymbol{\Sigma},$$

这里 $\delta_{\alpha\beta}$ 是 Kronecker 指标, 即 $\delta_{\alpha\beta}=\begin{cases}0, & \alpha\neq\beta, \\ 1, & \alpha=\beta.\end{cases}$ 因为因变量间这种联系的存在, 不宜使用(7.49)的记法, 而整体地记为

$$\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{\varepsilon}. \tag{7.50}$$

并附加假定

$$E\boldsymbol{\varepsilon}=0,\ \boldsymbol{\varepsilon}\ \text{各行不相关有共同协方差阵}\ \boldsymbol{\Sigma}>0. \tag{7.51}$$

在本书中总假定多元线性模型(7.49)满足假定(7.50), 以后再提到时常常不予补充说明. 很显然, 这正是一元的一般线性模型 $\boldsymbol{y}\sim(\boldsymbol{X\beta},\ \sigma^2\boldsymbol{I})$到多元情形的自然推广, 称为**一般多元线性模型**, 这里的多元是指因变量是多个的, 有时为明确起见也称为多指标的. 在第六章的回归分析中, 我们也提到过多元线性回归, 那时是针对多个自变量而言, 今后在谈到多元线性回归时可能指的是多指标情形, 相信在出现具体模型时, 不致引起混淆, 我们避免给习惯上已经形成的名词以另外的称谓.

在一元线性模型中参数是向量 $\boldsymbol{\beta}$ 和数 $\sigma^2>0$, 而在多元情形,

参数是矩阵 $\boldsymbol{B}$ 和协方差阵 $\boldsymbol{\Sigma}>0$. 这时，参数的统计推断，基本上都是一元情形的平行推广. 在估计问题中，这种平行推广不存在任何实质性困难，但在涉及假设检验和区间估计时，由于多元统计量分布的复杂性，有些困难就很不容易解决，但就方法的形式而言，仍然可以说是一致的.

多元线性统计推断与一元情形的极其相似可用这两种模型实质上的一致来说明. 事实上我们可以设想将观察值矩阵 $\boldsymbol{Y}$ 按它的行序将其元素逐个地记成一列，设为

$$\boldsymbol{y}=(y_{11}\cdots y_{1p}y_{21}\cdots y_{2p}\cdots y_{n1}\cdots y_{np})',$$

且将 $\boldsymbol{B}$, $\boldsymbol{\varepsilon}$ 也按上法处理记成

$$\boldsymbol{B}=(\beta_{11}\cdots\beta_{1p}\beta_{21}\cdots\beta_{2p}\cdots\beta_{q1}\cdots\beta_{qp})',$$

$$\boldsymbol{\varepsilon}=(\varepsilon_{11}\cdots\varepsilon_{1p}\varepsilon_{21}\cdots\varepsilon_{2p}\cdots\varepsilon_{n1}\cdots\varepsilon_{np})',$$

于是模型(7.50)可改写成

$$\boldsymbol{y}=\begin{pmatrix} x_{11} & 0 & & x_{1q} & 0 \\ & \ddots & \cdots & & \ddots \\ 0 & x_{11} & & 0 & x_{1q} \\ \vdots & & & \vdots & \\ x_{n1} & 0 & & x_{nq} & 0 \\ & \ddots & \cdots & & \ddots \\ 0 & x_{n1} & & 0 & x_{nq} \end{pmatrix}\boldsymbol{B}+\boldsymbol{\varepsilon}, \qquad (7.52)$$

条件(7.51)亦随之变为

$$\boldsymbol{E\varepsilon}=0,\ \mathrm{Cov}\boldsymbol{\varepsilon}=\begin{pmatrix} \boldsymbol{\Sigma} & & & 0 \\ & \boldsymbol{\Sigma} & & \\ & & \ddots & \\ 0 & & & \boldsymbol{\Sigma} \end{pmatrix}. \qquad (7.53)$$

虽然满足(7.53)的一元线性模型(7.52)与我们在§6.5中所讨论的一般线性模型有所不同，但不过是稍加推广而已，毕竟还是线性模型. 它是更广泛的一元线性模型的特例. 所以说多元线性模型的参数统计推断问题，本质上与一元的相同，可在一元模型中进行，由于我们在第六章中并未讨论如(7.52)和(7.53)的模型，并且为了保持多元统计推断问题的原来面目，我们的讨论将对模型

(7.50)进行，对于(7.52)就不予深入展开了。

(二)参数估计及其分布

现在对满足假定(7.51)的多元线性模型(7.50)考虑参数估计问题.

与一元情形相仿，如果不对设计矩阵的秩加以限制，就存在可估性问题. 但可估性问题取决于设计阵 $\boldsymbol{X}$，故在多元情形与一元情形完全类同，就不再予以赘述. 凡要用到的事实，就直接引用§6.5(一)中结果. 为了讨论方便，在大多数情形下，我们只考虑 $\boldsymbol{X}$ 满列秩情形，即设 $\mathrm{rk}\,\boldsymbol{X}=q$，并将可能出现的 $\boldsymbol{X}$ 的第一列全为1的情形，除非特别说明，一般也包括在内.

对于参数矩阵 $\boldsymbol{B}$ 的估计，可用一元情形下最小二乘法的推广. 在一元情形，设 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的估计，则讨论剩余(残差) $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$ 的平方和，即 $\hat{\varepsilon}'\hat{\varepsilon}$，使之极小化，所得估计量就是最小二乘估计，现在的问题是：设 $\hat{\boldsymbol{B}}$ 是 $\boldsymbol{B}$ 的估计，则剩余(残差) $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{B}}$ 是一个 $n\times p$ 阶矩阵，如果也要讨论它的"平方和极小"，究竟在什么意义下进行呢？相仿可取剩余"乘积和"矩阵为 $\hat{\varepsilon}'\hat{\varepsilon}$，称它为**残积阵**，而它的"极小值"可在非负定意义下去理解，即若 $\hat{\boldsymbol{B}}$ 是我们欲求的优良估计，我们希望它能满足

$$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})-(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{B}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{B}})\geqslant 0\text{(非负定阵)}$$

或记为

$$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\geqslant(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{B}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{B}}),\text{对一切 }\boldsymbol{B}. \tag{7.54}$$

由于当 $\boldsymbol{C}\geqslant\boldsymbol{D}\geqslant 0$ 时，必有它们的从大到小排列的顺序特征值，$\lambda_i(\boldsymbol{C})$，$\lambda_i(\boldsymbol{D})$满足 $\lambda_i(\boldsymbol{C})\geqslant\lambda_i(\boldsymbol{D})$，$i=1,\cdots,r$，这里 $r=\mathrm{rk}D$. 因此有 $\prod_{i=1}^{r}\lambda_i(\boldsymbol{C})\geqslant\prod_{i=1}^{r}\lambda_i(\boldsymbol{D})$，且有 $\mathrm{tr}\,\boldsymbol{C}\geqslant\mathrm{tr}\,\boldsymbol{D}$，等等性质，故(7.54)是一个很强的结果. 当 $p=1$，它就使残差平方和极小. 作为一元情形的推广，我们仍称满足(7.54)的估计为 $\boldsymbol{B}$ 的**最小二乘估计**.

定理 7.11

对于模型(7.50)(这里是否满足假定(7.51)是无关紧要的)，满足正规方程

$$X'XB=X'Y \tag{7.55}$$

的 $\hat{B}$ 是 B 的最小二乘估计，且对 $\hat{B}$ 而言，残积阵是

$$R_0=Y'P_{X^\perp}Y \tag{7.56}$$

证 用平方和分解法得

$$\begin{aligned}(Y-XB)'(Y-XB)&=(Y-X\hat{B})'(Y-X\hat{B})\\&\quad+(X\hat{B}-XB)'(X\hat{B}-XB)+(Y-X\hat{B})'(X\hat{B}-XB)\\&\quad+(X\hat{B}-XB)'(Y-X\hat{B}).\end{aligned}$$

由于 $\hat{B}$ 满足正规方程(7.55)，知上式右边后两项为零，得

$$\begin{aligned}&(Y-XB)'(Y-XB)-(Y-X\hat{B})'(Y-X\hat{B})\\&\quad=(X\hat{B}-XB)'(X\hat{B}-XB)\geqslant 0.\end{aligned}$$

如果上式不等号改为等号，充要条件是

$$X(\hat{B}-B)=0,$$

但由此可推出 $X'X(\hat{B}-B)=0$，由 $X'X$ 可逆得 $B=\hat{B}$. 故若另一估计的残积阵与 $\hat{B}$ 的残积阵相同，此估计必为 $\hat{B}$.

由于 $\hat{B}=(X'X)^{-1}X'Y$，即得

$$(Y-X\hat{B})'(Y-X\hat{B})=Y'(I-X(X'X)^{-1}X')Y=Y'P_{X^\perp}Y.$$

记此残积阵为 R_0，定理证毕.

注记：为了使 $\hat{B}$ 是 B 的最小二乘估计，$\hat{B}$ 满足(7.55)不仅是充分的，而且是必要的，必要性的证明留作练习.

与一元情形的定理 6.5 相仿可得

定理 7.12

在模型(7.50)下，参数的线性函数 $\mathrm{tr}\,C'B$ 的一切线性无偏估计 $\mathrm{tr}\,D'Y$ 中，$\mathrm{tr}\,C'\hat{B}$ 是唯一的极小方差估计，即 $\mathrm{tr}\,C'\hat{B}$ 是 $\mathrm{tr}\,C'B$ 的 BLUE. 其中 $\hat{B}=(X'X)^{-1}X'Y$ 是 B 的最小二乘估计.

证 因 $E\,\mathrm{tr}\,D'Y=\mathrm{tr}\,C'B$ 对一切 B 成立，知 $\mathrm{tr}\,D'XB=\mathrm{tr}\,C'B$ 或 $\mathrm{tr}(D'X-C')B=0$ 对一切 B 成立，故有 $D'X=C'$. 为无偏性的充要条件. 易见 $\hat{B}$ 是 B 的无偏估计. 注意到

$$\boldsymbol{X}\hat{\boldsymbol{B}}=\boldsymbol{P}_X\boldsymbol{Y} \tag{7.57}$$

我们有 $\operatorname{tr}\boldsymbol{C}'\hat{\boldsymbol{B}}=\operatorname{tr}\boldsymbol{D}'\boldsymbol{X}\hat{\boldsymbol{B}}=\operatorname{tr}\boldsymbol{D}'\boldsymbol{P}_X\boldsymbol{Y}$. 对于任意的 $\operatorname{tr}\boldsymbol{A}'\boldsymbol{Y}$, 我们有

$$D(\operatorname{tr}\boldsymbol{A}'\boldsymbol{Y})=D(\operatorname{tr}\boldsymbol{Y}\boldsymbol{A}')=D\left(\sum_{\alpha=1}^{n}\boldsymbol{y}_\alpha'\boldsymbol{a}_\alpha\right)=\sum_{\alpha=1}^{n}\boldsymbol{a}_\alpha'\boldsymbol{\Sigma}\boldsymbol{a}_\alpha$$
$$=\operatorname{tr}\boldsymbol{\Sigma}\boldsymbol{A}'\boldsymbol{A}.\quad (\boldsymbol{A}'=(\boldsymbol{a}_1\cdots\boldsymbol{a}_n))$$

因此 $D(\operatorname{tr}\boldsymbol{C}'\hat{\boldsymbol{B}})=D(\operatorname{tr}\boldsymbol{D}'\boldsymbol{P}_X\boldsymbol{Y})=\operatorname{tr}\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{P}_X\boldsymbol{D}\leqslant\operatorname{tr}\boldsymbol{\Sigma}\boldsymbol{D}'\boldsymbol{D}$

$$=D(\operatorname{tr}\boldsymbol{D}'\boldsymbol{Y}).$$

且上式相等的充要条件是 $\operatorname{tr}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{D}'(\boldsymbol{I}-\boldsymbol{P}_X)\boldsymbol{D}\boldsymbol{\Sigma}^{\frac{1}{2}}=0$, 即 $(\boldsymbol{I}-\boldsymbol{P}_X)\cdot\boldsymbol{D}\boldsymbol{\Sigma}^{\frac{1}{2}}=0$, 或 $\boldsymbol{D}=\boldsymbol{P}_X\boldsymbol{D}$. 于是有

$$\boldsymbol{D}'\boldsymbol{Y}=\boldsymbol{D}'\boldsymbol{P}_X\boldsymbol{Y}=\boldsymbol{D}'\boldsymbol{X}\hat{\boldsymbol{B}}=\boldsymbol{C}'\hat{\boldsymbol{B}}.$$ 证毕.

下面在正态性假定下进行讨论. 由 §7.1(二)的记号, 设

$$\varepsilon\sim N_{n,p}(0,\ I,\ \Sigma). \tag{7.58}$$

称满足(7.58)的模型(7.50)为正态模型.

定理 7.13

设正态模型(7.50), $\hat{\boldsymbol{B}}$ 是 $\boldsymbol{B}$ 的最小二乘估计, $\boldsymbol{R}_0=\boldsymbol{Y}'\boldsymbol{P}_{X^\perp}\boldsymbol{Y}$ 是残积阵, 记 $\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\boldsymbol{R}_0$, 则有 $(\hat{\boldsymbol{B}},\ \hat{\boldsymbol{\Sigma}})$ 是 $(\boldsymbol{B},\ \boldsymbol{\Sigma})$ 的极大似然估计.

证 这时似然函数为

$$\boldsymbol{L}(\boldsymbol{Y},\ \boldsymbol{B},\ \boldsymbol{\Sigma})$$
$$=C\cdot(\det\boldsymbol{\Sigma})^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{n}(\boldsymbol{y}_\alpha-\boldsymbol{B}'\boldsymbol{x}_\alpha)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_\alpha-\boldsymbol{B}'\boldsymbol{x}_\alpha)\right\}$$

其中 $\boldsymbol{x}_\alpha'$ 是 X 的第 α 行. 注意到

$$\sum_{\alpha=1}^{n}(\boldsymbol{y}_\alpha-\boldsymbol{B}'\boldsymbol{x}_\alpha)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_\alpha-\boldsymbol{B}'\boldsymbol{x}_\alpha)$$
$$=\operatorname{tr}\boldsymbol{\Sigma}^{-1}\sum_{\alpha=1}^{n}(\boldsymbol{y}_\alpha-\boldsymbol{B}'\boldsymbol{x}_\alpha)(\boldsymbol{y}_\alpha-\boldsymbol{B}'\boldsymbol{x}_\alpha)'$$
$$=\operatorname{tr}\Sigma^{-1}[(Y-X\hat{\boldsymbol{B}})'(Y-X\hat{\boldsymbol{B}})+(X\hat{\boldsymbol{B}}-XB)'(X\hat{\boldsymbol{B}}-XB)]$$
$$\geqslant\operatorname{tr}\Sigma^{-1}Y'P_{X^\perp}Y=\operatorname{tr}\Sigma^{-1}R_0$$

且上式等号成立的充要条件是 $B=\hat{\boldsymbol{B}}$. 以下证明与定理 7.9 的证明完全相同. 得证.

注记: 与定理 7.8 相仿, 我们可证 $(\hat{\boldsymbol{B}},\ R_0/n-q)$ 还是 $(B,\ \Sigma)$

的极小方差无偏估计. 但方法完全类似就不予赘述了. 至于 $R_0/(n-q)$是 Σ 的无偏估计,可在不假定正态性的一般条件下验证,事实上因 $\boldsymbol{P}_{X^\perp}$ 是秩为 $n-q$ 的正投影阵,必存在正交阵 $\boldsymbol{U}$ 使 $\boldsymbol{U}'\boldsymbol{P}_{X^\perp}\boldsymbol{U}=\begin{pmatrix}\boldsymbol{I}_{n-q} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}$, 于是

$$\boldsymbol{R}_0=\boldsymbol{Y}'\boldsymbol{U}\boldsymbol{U}'\boldsymbol{P}_{X^\perp}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{Y}=\boldsymbol{Y}'\boldsymbol{U}\begin{pmatrix}\boldsymbol{I}_{n-q} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{U}'\boldsymbol{Y},$$

注意到 $\boldsymbol{Z}\triangleq\boldsymbol{U}'\boldsymbol{Y}\sim N_{np}(\boldsymbol{0},\ \boldsymbol{I},\ \boldsymbol{\Sigma})$, 记 $\boldsymbol{Z}'=(\boldsymbol{Z}_1\cdots\boldsymbol{Z}_n)$有 $\boldsymbol{R}_0=\sum_{\alpha=1}^{n-q}\boldsymbol{Z}_\alpha\boldsymbol{Z}'_\alpha$, 因此得

$$E\boldsymbol{R}_0=\sum_{\alpha=1}^{n-q}E\boldsymbol{Z}_\alpha\boldsymbol{Z}'_\alpha=(n-q)\boldsymbol{\Sigma} \tag{7.59}$$

关于 $\hat{\boldsymbol{B}}$ 和 $\boldsymbol{R}_0$ 的分布,我们有

定理 7.14

设正态模型(7.50), $\hat{\boldsymbol{B}}$, $\boldsymbol{R}_0$ 分别是 $\boldsymbol{B}$ 的最小二乘估计和残积阵,则有

1° $\hat{\boldsymbol{B}}\sim N_{qp}(\boldsymbol{B},\ (\boldsymbol{X}'\boldsymbol{X})^{-1},\ \boldsymbol{\Sigma})$;

2° $\boldsymbol{R}_0\sim W_p(n-q,\ \boldsymbol{\Sigma})$;

3° $\hat{\boldsymbol{B}}$ 与 $\boldsymbol{R}_0$ 独立.

证 $\hat{\boldsymbol{B}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, 由定理 7.1 立得 1°. 注意到 $\boldsymbol{B}'\boldsymbol{X}'\boldsymbol{P}_{X^\perp}\boldsymbol{X}\boldsymbol{B}=0$, 由定理 7.2、7.3. 1° 得 2°. 再由$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}_{X^\perp}=0$, 根据定理 7.3.2° 得 3°.

现在我们把回归和预测的概念推广到多元情形中来, 设因变量 $Y_1,\ \cdots,\ Y_p$ 对自变量 $X_1,\ \cdots X_q$ 的统计依赖表现为: 当 $X_1,\ \cdots,\ X_q$ 给定时, 有如下的线性模型:

$$(Y_1\cdots Y_p)=(1x_1\cdots x_q)\boldsymbol{B}+(\varepsilon_1\cdots\varepsilon_p) \tag{7.60}$$

其中

$$\boldsymbol{B}=\begin{pmatrix}\beta_{01} & \cdots & \beta_{0p}\\ \beta_{11} & \cdots & \beta_{1p}\\ \vdots & & \vdots\\ \beta_{q1} & \cdots & \beta_{qp}\end{pmatrix}$$

称 β_{ij} 是第 i 个自变量 X_i 对第 j 个因变量 Y_j 的效应, (7.60)就

是总体回归模型.

由§7.1(一)性质5(看(7.8)式)可以看出，当$(Y_1, \cdots, Y_p, X_1, \cdots, X_q)' \triangleq \boldsymbol{Z} \sim N_{p+q}(\boldsymbol{\nu}, \boldsymbol{V})$，记

$$\boldsymbol{\nu}=\begin{pmatrix} E\boldsymbol{Y} \\ E\boldsymbol{X} \end{pmatrix}, \quad \boldsymbol{V}=\begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}, \quad \boldsymbol{V}>0,$$

则当$X_1, \cdots, X_q$给定时，$\boldsymbol{Y}=(Y_1, \cdots, Y_p)'$的条件分布

$$\boldsymbol{Y}|\boldsymbol{X} \sim N_p(E\boldsymbol{Y}+\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X}-E\boldsymbol{X}), \ \boldsymbol{\Sigma}_{YY}-\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}).$$

可见总体回归函数是

$$E\boldsymbol{Y}+\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X}-E\boldsymbol{X}). \tag{7.61}$$

不难看出，以回归方程

$$\hat{\boldsymbol{Y}}=E\boldsymbol{Y}+\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{X}-E\boldsymbol{X}) \tag{7.62}$$

作Y的预测，预测误差的协方差阵是

$$\boldsymbol{\Sigma}_{YY}-\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}. \tag{7.63}$$

这些都是一元情形的自然推广.

相仿，由最小二乘估计得回归系数的估计

$$\begin{aligned}
\hat{\boldsymbol{B}} &= ((\boldsymbol{1} \vdots \boldsymbol{X})'(\boldsymbol{1} \vdots \boldsymbol{X}))^{-1}(\boldsymbol{1} \vdots \boldsymbol{X})'\boldsymbol{Y} \\
&= \begin{pmatrix} n & \boldsymbol{1}'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{1} & \boldsymbol{X}'\boldsymbol{X} \end{pmatrix}^{-1} (\boldsymbol{1} \vdots \boldsymbol{X})'\boldsymbol{Y} \\
&= \begin{pmatrix} \dfrac{1}{n}+\dfrac{\boldsymbol{1}'\boldsymbol{X}}{n}(\boldsymbol{X}'\boldsymbol{P}_{1^\perp}\boldsymbol{X})^{-1}\dfrac{\boldsymbol{X}'\boldsymbol{1}}{n} & \dfrac{-\boldsymbol{1}'\boldsymbol{X}}{n}(\boldsymbol{X}'\boldsymbol{P}_{1^\perp}\boldsymbol{X})^{-1} \\ -(\boldsymbol{X}'\boldsymbol{P}_{1^\perp}\boldsymbol{X})^{-1}\dfrac{\boldsymbol{X}'\boldsymbol{1}}{n} & (\boldsymbol{X}'\boldsymbol{P}_{1^\perp}\boldsymbol{X})^{-1} \end{pmatrix} \\
&\quad \times \begin{pmatrix} \boldsymbol{1}'\boldsymbol{Y} \\ \boldsymbol{X}'\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{1}'\boldsymbol{Y}/n-\boldsymbol{1}'\boldsymbol{X}/n(\boldsymbol{X}'\boldsymbol{P}_{1^\perp}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{P}_{1^\perp}\boldsymbol{Y}) \\ (\boldsymbol{X}'\boldsymbol{P}_{1^\perp}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{P}_{1^\perp}\boldsymbol{Y}) \end{pmatrix}.
\end{aligned} \tag{7.64}$$

这与总体回归情形以样本矩替代总体矩所得结果完全一致. 并且与一元情形也十分相仿.

(三)线性假设检验

先讨论回归模型(7.60)的假设检验. 这时要检验的假设是

$$H_0: \boldsymbol{B}_1 \triangleq \begin{pmatrix} \beta_{11} & \cdots & \beta_{1p} \\ \vdots & & \vdots \\ \beta_{q1} & \cdots & \beta_{qp} \end{pmatrix} = 0.$$

对于正态模型，可由似然比导出检验统计量．令

$$\lambda = \frac{\max\{L(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{\Sigma}): \boldsymbol{B}, \boldsymbol{\Sigma}>0\}}{\max\{L(\boldsymbol{Y}, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}): \boldsymbol{\beta}_0, \boldsymbol{\Sigma}>0\}} \triangleq \frac{M}{M_H}.$$

其中 $\boldsymbol{\beta}_0$ 是 $\boldsymbol{B}$ 的第一个行向量．由定理 7.13 已知 $M = C[\det(\boldsymbol{R}_0/n)]^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\right\}$．而 H_0 成立时

$$L(\boldsymbol{Y}, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}) = C\cdot(\det\boldsymbol{\Sigma})^{-\frac{n}{2}}\cdot\exp\left\{-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1} \times \sum_{\alpha=1}^{n}(\boldsymbol{y}_\alpha-\boldsymbol{\beta}_0)(\boldsymbol{y}_\alpha-\boldsymbol{\beta}_0)'\right\}$$

注意到 $\sum_{\alpha=1}^{n}(\boldsymbol{y}_\alpha-\boldsymbol{\beta}_0)(\boldsymbol{y}_\alpha-\boldsymbol{\beta}_0)' = \sum_{\alpha=1}^{n}(\boldsymbol{y}_\alpha-\bar{\boldsymbol{y}})(\boldsymbol{y}_\alpha-\bar{\boldsymbol{y}}) + n(\bar{\boldsymbol{y}}-\boldsymbol{\beta}_0)(\bar{\boldsymbol{y}}-\boldsymbol{\beta}_0)' = \boldsymbol{Y}'\boldsymbol{P}_{1^\perp}\boldsymbol{Y} + n(\bar{\boldsymbol{y}}-\boldsymbol{\beta}_0)(\bar{\boldsymbol{y}}-\boldsymbol{\beta}_0)'$,

得
$$M_H = C\cdot[\det(\boldsymbol{Y}'\boldsymbol{P}_{1^\perp}\boldsymbol{Y}/n)]^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}\right\}.$$

记 $\boldsymbol{Y}'\boldsymbol{P}_{1^\perp}\boldsymbol{Y} = \boldsymbol{R}_1$．有 $\lambda = (\det\boldsymbol{R}_0/\det\boldsymbol{R}_1)^{-\frac{n}{2}}$，它是 $\Lambda \triangleq (\det\boldsymbol{R}_0/\det\boldsymbol{R}_1)$的严降函数，因此检验的否定域取$\{\Lambda \leqslant C_\alpha\}$．由于 $\boldsymbol{R}_0 \sim W_p(n-q-1, \boldsymbol{\Sigma})$[1]，当 H_0 成立时，$\boldsymbol{R}_1 \sim W_p(n-1, \boldsymbol{\Sigma})$，得 $\boldsymbol{R}_1-\boldsymbol{R}_0$ 与 $\boldsymbol{R}_0$ 独立，且 $\boldsymbol{R}_1-\boldsymbol{R}_0 \sim W_p(q, \boldsymbol{\Sigma})$，故有

$$\Lambda \sim \Lambda(p, n-q-1, q).$$

当检验水平 α 给定时，可查表得临界值 C_α．

当 H_0 被拒绝，我们接受回归模型(7.60)，与一元情形相仿，尚需检验自变量 X_i 的效应的显著性，这时要检验的假设是

$$H_{0i}: \beta_{i1} = \cdots = \beta_{ip} = 0.$$

仿上用似然比检验，这时容易验证，当 H_{0i} 成立时，作为受约束模型求 B 的最小二乘估计，其残积阵为 $\boldsymbol{R}_{1i} = \boldsymbol{Y}'\boldsymbol{P}_{\mu_i^\perp}\boldsymbol{Y}$，其中 μ_i 是在

1) 欲使 R_0 概率为 1 地非奇异，要求 $n-q-1 \geqslant p$．

$(\mathbf{1X})$中划去 $\boldsymbol{X}$ 的第 i 列后所张成的线性空间，于是有 $\boldsymbol{R}_{1i}\sim W_p(n-q,\ \boldsymbol{\Sigma},\ \boldsymbol{\Delta})$. 当假设成立时，我们有 $\boldsymbol{R}_{1i}\sim W_p(n-q,\ \boldsymbol{\Sigma})$, $\boldsymbol{R}_{1i}-\boldsymbol{R}_0\sim W_p(1,\ \boldsymbol{\Sigma})$. 于是有

$$\Lambda=\frac{\det \boldsymbol{R}_0}{\det \boldsymbol{R}_{1i}}\sim\Lambda(p,\ n-q-1,\ 1)$$

由表 7.1 知，令 $F=\frac{1-\Lambda}{\Lambda}\ \frac{n-q-p}{p}$，当 H_{0i} 成立时有

$$\frac{1-\Lambda}{\Lambda}\ \frac{n-q-p}{p}\sim F_{p,n-p-q},$$

故得水平为 α 的检验的拒绝域是

$$\left\{\Lambda^{-1}\geqslant 1+\frac{p}{n-p-q}\cdot F_{p,n-p-q,\alpha}\right\}.$$

注意，在进行这一检验时，我们必须限制 $n>p+q$.

在选择变量问题上，多元线性模型的情形比一元复杂，因为若自变量 X_i 对整个因变量集的效应即使是显著的(否定 H_{0i})，但不见得它对每个因变量的效应都显著. 从而可提出检验参数矩阵 B 的列向量是否为零的问题，如果为零意味着相应的因变量并不统计依赖于现有自变量因子，这时要考虑选进其它因子，才有可能预报这一因变量. 另外，一个自变量虽然对整个 Y 的效应并不显著，但对某个因变量仍有可能是显著的. 总之，这时的自变量选择问题变得错综复杂，如何定出合理的选择标准，给出相应的选择方法，是一个很有实际意义的问题. 但这里无法给予深入讨论.

现在我们来讨论一般线性假设

$$H_0:\ \boldsymbol{HBG}=\boldsymbol{O}\ \text{的检验.}$$

这里 $\boldsymbol{H}$ 是 $k\times q$ 阶阵，$\boldsymbol{G}$ 是 $p\times r$ 阶阵，且有 $\operatorname{rk}\boldsymbol{H}=k$, $\operatorname{rk}\boldsymbol{G}=r$. 这类假设表现了多元线性模型参数统计推断问题的特点. 如单纯考虑 $\boldsymbol{HB}=\mathbf{0}$ 型的假设，实际上是对 $\boldsymbol{B}$ 的列向量所张成的空间的约束，它完全可以从一元检验问题平行地推广过来，但当假设为 $\boldsymbol{BG}=\mathbf{0}$ 型，就是约束了 $\boldsymbol{B}$ 的行向量所张成的线性空间，这类检验问题在一元统计分析中是不存在的，因而需要寻求特别的方法去检验. 然而后一类型的检验问题可以通过适当的变换化归为前一类

型处理.

先讨论检验一般线性假设

$$H_0:\quad \boldsymbol{HB}=\mathbf{0}.$$

我们只想说明如何将一元线性假设检验的方法平行推广过来，而不详述推导检验统计量的全过程. 注意到最小二乘估计所给出的估计量实际上可按列分别给出，即若把多元线性模型看成如(7.49)的 p 个一元线性模型，分别求 β_j 的最小二乘估计，然后将 $\hat{\boldsymbol{\beta}}_j$ 合成一个矩阵，恰好就是多元线性模型中 $\boldsymbol{B}$ 的最小二乘估计 $\hat{\boldsymbol{B}}$. 考虑 $\boldsymbol{HB}=\mathbf{0}$ 是一个约束，实际上也是 $\boldsymbol{H\beta}_1=\cdots=\boldsymbol{H\beta}_p=\mathbf{0}$, 所以求多元线性模型(7.50)在约束 $\boldsymbol{HB}=\mathbf{0}$ 下的最小二乘估计 $\hat{\boldsymbol{B}}_H$, 也可以分别考虑(7.49)中各个一元线性模型在约束 $\boldsymbol{H\beta}_j=\mathbf{0}$ 下的最小二乘估计，故有 $\boldsymbol{X}\hat{\boldsymbol{B}}_H=\boldsymbol{P}_{XQ}\boldsymbol{Y}$, 这里 $\boldsymbol{Q}$ 是一个 q 阶方阵，使得 $\mu(\boldsymbol{Q})=N(\boldsymbol{H})$. 实际上可取

$$\boldsymbol{Q}=\boldsymbol{I}-\boldsymbol{H}'(\boldsymbol{HH}')^{-1}\boldsymbol{H}\triangleq P_{H'^{\perp}},\tag{7.65}$$

它是到 $\mu(\boldsymbol{H}')$ 的正交补空间的投影阵.(参看§6.5(二))记 $\hat{\boldsymbol{B}}_H$ 的残积阵为 $\boldsymbol{R}_H$, 有

$$\boldsymbol{R}_H=\boldsymbol{Y}'\boldsymbol{P}_{(XQ)^{\perp}}\boldsymbol{Y}\tag{7.66}$$

从而仿前面的推导可得似然比 $\lambda=M/M_H$ 是

$$\Lambda=\frac{\det\boldsymbol{R}_0}{\det\boldsymbol{R}_H}$$

的严降函数. 因此，检验 $\boldsymbol{HB}=\mathbf{0}$ 的拒绝域是 $\{\Lambda\leqslant C_\alpha\}$. 根据§6.5(二)中(6.73)式，知

$$\operatorname{rk}\boldsymbol{P}_{(XQ)^{\perp}}=n-[\operatorname{rk}(\boldsymbol{X}'\,|\,\boldsymbol{H}')-\operatorname{rk}\boldsymbol{H}']\triangleq n-s.$$

其中 $s\triangleq\operatorname{rk}(\boldsymbol{X}'\,|\,\boldsymbol{H}')-\operatorname{rk}\boldsymbol{H}'$. 因此，当 H_0 成立时，

$$\Lambda\sim\Lambda(p,\ n-q,\ q-s).$$

它只有在特殊情况下等价于 F 检验，一般就要查 Λ-分布表. 从而确定水平为 α 的检验的临界值.

为了检验假设 H_0: $\boldsymbol{HBG}=\mathbf{0}$, 我们令

$$\boldsymbol{Z}=\boldsymbol{YG},\ \boldsymbol{\Theta}=\boldsymbol{BG},$$

模型(7.50)就变为

$$\boldsymbol{Z}=\boldsymbol{X\Theta}+\varepsilon\boldsymbol{G} \tag{7.67}$$

$$\varepsilon\boldsymbol{G}\sim N_{nr}(\boldsymbol{0},\ \boldsymbol{I},\ \boldsymbol{G}'\boldsymbol{\Sigma}\boldsymbol{G}) \tag{7.68}$$

假设 H_0: $\boldsymbol{HBG}=\boldsymbol{0}$ 就变为

$$\widetilde{H}_0:\ \boldsymbol{H\Theta}=\boldsymbol{0}$$

从而可仿上面讨论得到在模型(7.67)下检验 $\widetilde{H}_0$ 的检验统计量是

$$\Lambda=\det(\boldsymbol{Z}'\boldsymbol{P}_{X^\perp}\boldsymbol{Z})/\det(\boldsymbol{Z}'\boldsymbol{P}_{(XQ)^\perp}\boldsymbol{Z})$$

其中 Q 如(7.65)定义. 由原模型数据表出就是

$$\Lambda=\det(\boldsymbol{G}'\boldsymbol{Y}'\boldsymbol{P}_{X^\perp}\boldsymbol{YG})/\det(\boldsymbol{G}'\boldsymbol{Y}'\boldsymbol{P}_{(XQ)^\perp}\boldsymbol{YG}).$$

当假设 H_0 成立,有

$$\Lambda\sim\Lambda(r,\ n-q,\ q-s)$$

其中 $s=\mathrm{rk}(\boldsymbol{X}'\,|\,\boldsymbol{H}')-\mathrm{rk}\,\boldsymbol{H}'$, 其余同上.

在讨论假设检验时, 有一个常用的方法是将模型化为某种典型的型式,称为法式. 在法式下检验问题变得十分明确,便于进行讨论. 这种方法是由许宝騄教授在1941年提出的, 它在讨论 F-检验的优良性时起了明显的作用.

最后我们来讨论广义方差分析. 在一元统计分析中, 对于一些狭义的方差分析模型(一向分类、两向分类模型等),用方差分析方法去检验因子效应的效著性有明显的方便之处, 它可以通过简单的和式与乘积的计算去实现,在多元统计分析中,它的自然的推广就是广义方差分析, 实际上也就是将观察值矩阵的"乘积和"阵 $\boldsymbol{Y}'\boldsymbol{Y}$ 进行分解,如

$$\boldsymbol{Y}'\boldsymbol{Y}=\sum_{i=1}^{k}\boldsymbol{W}_i$$

这里 $\boldsymbol{W}_i$ 是一些相互独立的遵从 Wishart 分布的随机矩阵, 根据 Cochran 定理到矩阵情形的推广(看定理7.5), 利用 Wilks 统计量 Λ, 可按分解式进行一些有关的检验.

例如考虑两向分类模型

$$\boldsymbol{y}_{ij}=\boldsymbol{\theta}_0+\boldsymbol{\beta}_i+\boldsymbol{\gamma}_j+\varepsilon_{ij},\quad i=1,\ \cdots,\ r;\ j=1,\ \cdots,\ c. \tag{7.69}$$

其中 $\boldsymbol{y}_{ij}$, $\boldsymbol{\theta}_0$, $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}_j$, ε_{ij} 都是 p 维向量. 仿照对(6.50)的方差分析,我们易得分解式

$$\boldsymbol{Y}'\boldsymbol{Y}=\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}'+\sum_i C(\bar{\boldsymbol{y}}_{i\cdot}-\bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_{i\cdot}-\bar{\boldsymbol{y}})'+\sum_j r(\bar{\boldsymbol{y}}_{\cdot j}-\bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_{\cdot j}-\bar{\boldsymbol{y}})'$$
$$+\sum_i\sum_j(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_{i\cdot}-\bar{\boldsymbol{y}}_{\cdot j}+\bar{\boldsymbol{y}})(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_{i\cdot}-\bar{\boldsymbol{y}}_{\cdot j}+\bar{\boldsymbol{y}})'.$$
$$\triangleq \boldsymbol{W}_0+\boldsymbol{W}_r+\boldsymbol{W}_c+\boldsymbol{W}_s$$

其中 $\bar{\boldsymbol{y}}=\frac{1}{rc}\sum_{i,j}\boldsymbol{y}_{ij}$, $\bar{\boldsymbol{y}}_{i\cdot}=\frac{1}{c}\sum_{j=1}^{c}\boldsymbol{y}_{ij}$, $\bar{\boldsymbol{y}}_{\cdot j}=\frac{1}{r}\sum_{i=1}^{r}\boldsymbol{y}_{ij}$. 欲检验假设

$$H_{01}:\ \boldsymbol{\beta}_1=\cdots=\boldsymbol{\beta}_r=\boldsymbol{0},$$
$$H_{02}:\ \boldsymbol{\gamma}_1=\cdots=\boldsymbol{\gamma}_c=\boldsymbol{0}.$$

检验 H_{01} 的统计量可取

$$\Lambda_1=\frac{\det \boldsymbol{W}_s}{\det(\boldsymbol{W}_s+\boldsymbol{W}_r)},$$

检验 H_{02} 的统计量可取

$$\Lambda_2=\frac{\det \boldsymbol{W}_s}{\det(\boldsymbol{W}_s+\boldsymbol{W}_c)}.$$

当零假设成立时，它们分别遵从 Λ-分布. 从而可仿前面的做法给出检验.

一般地说. 广义方差分析总可以由方差分析导出. 这是因为在 Wishart 矩阵和 χ^2 变量之间存在着内在联系. 欲给出多元线性模型中 $\boldsymbol{Y}'\boldsymbol{Y}$ 的分解式

$$\boldsymbol{Y}'\boldsymbol{Y}=\sum_{i=1}^{k}\boldsymbol{Y}'\boldsymbol{P}_{\mu_i}\boldsymbol{Y}. \tag{7.70}$$

则可考虑向量 $\boldsymbol{Ya}$ 的模型

$$\boldsymbol{Ya}=\boldsymbol{XBa}+\boldsymbol{\varepsilon a} \tag{7.71}$$

记 $\boldsymbol{Z}=\boldsymbol{Ya}$, 如有 $\boldsymbol{Z}'\boldsymbol{Z}$ 的分解式

$$\boldsymbol{Z}'\boldsymbol{Z}=\sum_{i=1}^{k}\boldsymbol{Z}'\boldsymbol{P}_{\mu_i}\boldsymbol{Z} \tag{7.72}$$

满足 $R^n=\mu_1\dot{+}\cdots\dot{+}\mu_k$, 这里 $\dot{+}$ 表示正交直和. 表示对(7.71)可用方差分析方法. 那么也就有与(7.72)相应的(7.70)式. 因为这里的关键是将空间作正交直和分解. 如果对(7.72)能作出明确的统计解释，则可将此解释搬用到(7.70)去，从而得到广义方差分析.

附注：**广义方差是方差概念的推广**，它一般指随机向量的协方差阵的行列式，也可定义为协方差阵的迹或最大特征值. 从而

可称样本协方差阵的行列式为样本广义方差，这里的广义方差分析就因此而得名.

§7.4　随机向量的互依性

在上一节的回归分析中，我们讨论了一组随机变量对另一组随机变量的统计依赖性. 本节将以随机向量的互依性作为研究对象，并从而得到一些实用的多元统计方法. 所谓互依性，就是把一个随机向量的各分量之间或两个随机向量之间的统计联系认为是相互的，而不是与自变量和因变量的关系相仿的依赖性. 在得到了一批样本观察值，但并不了解变量之间的关系时，从互依性的角度去分析这些观察值，似乎是更合理的.

研究互依性的目的，是找出随机向量或随机变量之间的主要联系. 从统计的角度看，就是要突出一组指标中起着主要作用的成分(可能是这组指标的某个函数)，从而可在实用中减少指标的个数(降低维数)，但不会造成明显的损失. 另一方面，观察值之间(变量之间)的相互依赖，有时会给统计处理和统计解释带来困难，因此，考虑对它们作一定的变换，使之相互独立，或作某种分解，使统计解释比较容易进行，在实践中也是有意义的.

下面我们分段讨论在互依性方面常用的几种统计方法：主成分分析，典型相关和因子分析.

(一)主成分分析

在对某个实际的多元统计分析问题进行讨论时，事先很可能不了解问题所涉及的多个指标之间的互依性，为了获取充分的信息从而作出较可靠的推断，我们往往选择许多个指标去进行观测，这些指标甚至多到十几个、几十个，然而，这些指标的意义究竟如何，是否可以剔除一些意义不大的指标或考虑引进更能说明问题的综合指标，无疑是值得探讨的. 例如在服装定型的研究中，有人对 256 个成年男人的体型，按 16 项指标进行了测量，经过主成分

分析，最后确定选择三个主成分作为定型的依据，从而使定型工作既简单易行，又有充分可靠的根据．又如要反映物价情况，对各种物价作全面调查固然可说明物价情况，但物价之间明显存在互依性，实际上选择几种主要商品的价格或得到某些综合指标，就可能足以反映物价的情况．有人对土壤标本进行研究，测了五项指标，经过主成分分析，发现其各项指标主要是由其中的两项所控制．这些情况都表明，在对数据进行统计分析之后，事先考虑观测的众多指标并不都是必要的，有时有大大压缩指标维数的可能，这样不但在经济上能节约开支减少观测，更重要的或许是使得主要矛盾突出，有利于解决实际问题．

主成分分析就是为了解决上述问题而引进的统计方法，它首先由 Hotelling 于 1933 年提出．从纯理论的角度看，这个方法是清晰易行的．它考虑对随机向量 $\boldsymbol{X}=(X_1, \cdots, X_p)'$ 作正交变换，令 $\boldsymbol{y}=\boldsymbol{U}'\boldsymbol{X}$，这里 $\boldsymbol{U}$ 是正交阵．一方面使 y 有较简单的协方差阵，如对角阵，于是 $\boldsymbol{y}$ 的各分量是不相关的．这时 y 的各个分量在整个变异中的作用就很容易衡量（下面将给出定义），这就使得我们有可能从 $\boldsymbol{y}$ 的分量中选择主要成分，剔除影响微弱的部分，也可以通过 $\boldsymbol{y}$ 的主成分的构造去分析原来的 $\boldsymbol{X}$ 的各个分量的作用．然而，在将主成分分析应用于实际问题时，量纲和尺度的问题会引起麻烦．如果有一些指标具有不同的量纲，要对这些指标求线性组合得综合指标，综合指标究竟反映什么呢？严格地说不同量纲的数是不能相加的．况且，即使形式上可以求和，尺度的选取将改变结果，而究竟该取怎样的尺度则因量纲不同而无统一标准．在主成分分析中解决这一问题的办法，是将每个指标值除以它的标准差，从而化成无量纲的纯数量．这似乎是一种可循的办法．可是，也可以提出如下问题：如 $\boldsymbol{X}=(X_1, \cdots, X_p)'$ 是对人体体型的 p 个特征的测量，量纲都是长度，譬如说身高是 150 厘米到 200 厘米，脚长是 15 厘米到 30 厘米，如此等等，把这类数据除以相应的标准差是否比不除标准差更合理呢？这很难从理论上给出回答，但结果却可能很不相同．如何去解释结果的差异呢？这看来是一

个棘手的问题．对这类问题的处理，单纯从理论上无法解决，恐怕要依靠对问题的实际知识．

下面，我们在假设$\boldsymbol{X}$的总体二阶矩已知的条件下，从数学的角度把这个问题探讨清楚．设$\mathrm{Cov}\,\boldsymbol{X}=\boldsymbol{\Sigma}$．那么，欲求正交阵$\boldsymbol{U}$使得$\boldsymbol{y}=\boldsymbol{U}'\boldsymbol{X}$的协方差矩阵为对角阵，且使主对角元按其次序由大到小排列，实际上正是矩阵代数中将$\Sigma\geqslant0$化为对角形的问题（见附录 A.4.1），即

$$\mathrm{Cov}\,\boldsymbol{y}=\boldsymbol{U}'\mathrm{Cov}\boldsymbol{X}\boldsymbol{U}=\boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U}=\boldsymbol{\Lambda},\tag{7.73}$$

$\boldsymbol{\Lambda}$是对角阵$\mathrm{diag}(\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_p)$，其中$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant0$．如果协方差阵$\boldsymbol{\Sigma}$的秩$\mathrm{rk}\,\boldsymbol{\Sigma}=r$，则有$\lambda_r>\lambda_{r+1}=0$．这里$\lambda_i$是$\boldsymbol{\Sigma}$的特征值，记$\boldsymbol{U}=(\boldsymbol{u}_1\cdots\boldsymbol{u}_r\boldsymbol{u}_{r+1}\cdots\boldsymbol{u}_p)$，则$\boldsymbol{u}_i$正是$\boldsymbol{\Sigma}$的相应于$\lambda_i$的特征向量，有$\boldsymbol{\Sigma}\boldsymbol{u}_i=\lambda_i\boldsymbol{u}_i,i=1,\ \cdots,\ p$．如果特征值中有一部分为零（在这里就是$\lambda_{r+1}=\cdots=\lambda p=0$），意味着$y_{r+1},\ \cdots,y_p$是没有变异的，这些分量就不必在统计中予以讨论．一般地说，$\boldsymbol{\Sigma}$不一定退化，但可能仍有一些特征值相对地很小，这表明$\boldsymbol{y}$在某些方向变异很小，可形象地认为象一块被压扁到几近平面的空间图形，其厚度可予忽略，并近似地当作平面图形来考虑．

定义 7.3

设$\boldsymbol{X}$是p维随机向量，$\boldsymbol{U}$是正交矩阵，使得$\boldsymbol{Y}=\boldsymbol{U}'\boldsymbol{X}$具有如(7.73)的协方差阵，则称$y_i$是（关于$\boldsymbol{X}$的）第$\boldsymbol{i}$个主成分，$i=1,\ \cdots,r$．称$\lambda_i/\sum_1^r\lambda_i$是第$i$个主成分的贡献率，称$\sum_1^k\lambda_i/\sum_1^r\lambda_i$是前$k$个主成分的累计贡献率．

由于$y_i=\boldsymbol{u}_i'\boldsymbol{X}$，所以一个主成分对应着$\boldsymbol{\Sigma}$的一个特征向量．$y_i$的方差$D(\boldsymbol{y}_i)=\boldsymbol{u}_i'\boldsymbol{\Sigma}\boldsymbol{u}_i=\lambda_i$反映了$\boldsymbol{y}_i$的变异，由于$\tilde{y}=(y_1,\ \cdots,\ y_r)'$的分量是互不相关的（当$\boldsymbol{X}$有正态分布，就是相互独立的），$\tilde{\boldsymbol{y}}$的变异由$\lambda_1,\ \cdots,\ \lambda_r$反映，贡献率的概念刻划了$y_i$的变异在整个变异中的地位．按照我们的定义，第一主成分的贡献率最大，并依次递减．如果前k个主成分的累计贡献率已接近 100%，那么，用这k个主成分作指标，比起用原来p个指标($\boldsymbol{X}$)，其反映实际问题

的能力已几乎一致，这就使问题可从 p 维空间降低到 k 维空间来讨论(一般有 $k<p$). 在实际应用中，当前 k 个主成分的累计贡献率达 85% 以上，就认为这 k 个主成分已足够反映原向量的变异.

我们还可从另一理论角度去解释 $\boldsymbol{X}$ 的信息为什么已经包含在关于 $\boldsymbol{X}$ 的 r 个主成分中. 现在考虑用 $\boldsymbol{X}$ 的 k 个线性函数 $\boldsymbol{b}_1'\boldsymbol{X}, \cdots, \boldsymbol{b}_k'\boldsymbol{X}$ 对 X_j 作线性预测，这里 $(\boldsymbol{b}_1'\boldsymbol{X}, \cdots, \boldsymbol{b}_k'\boldsymbol{X})'$ 是非退化的，即若记 $\boldsymbol{B}=(\boldsymbol{b}_1\cdots\boldsymbol{b}_k)$, 有 $\mathrm{Cov}\,\boldsymbol{B}'\boldsymbol{X}=\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B}>\boldsymbol{O}$. 要求预测的均方误差最小，记这个最小均方误差为

$$t_j(\boldsymbol{B}'\boldsymbol{X})=\min_{\boldsymbol{\beta}\in R^k} E(X_j-\boldsymbol{\beta}'\boldsymbol{B}\boldsymbol{X})^2, \quad j=1, \cdots, p.$$

我们有

定理 7.14

设 $\boldsymbol{u}_1'\boldsymbol{x}, \cdots, \boldsymbol{u}_k'\boldsymbol{x}$ 是关于 $\boldsymbol{X}$ 的前 k 个主成分，记 $\boldsymbol{U}_1=(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r)$(有 $\boldsymbol{U}=(\boldsymbol{U}_1 \vdots \boldsymbol{U}_2)$ 满足(7.73)). 则有

$$\sum_{j=1}^{p} t_j(\boldsymbol{U}_1'\boldsymbol{X})=\min_{\mathrm{rk}\boldsymbol{B}=k}\sum_{j=1}^{p} t_j(\boldsymbol{B}'\boldsymbol{X}) \tag{7.74}$$

证 由第 6 章 § 6.2 的定理 6.2 知，

$$t_j(\boldsymbol{B}'\boldsymbol{X})=D(X_j)-\mathrm{Cov}(X_j, \boldsymbol{B}'\boldsymbol{X})(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\mathrm{Cov}(X_j, \boldsymbol{B}'\boldsymbol{X})'.$$

因此
$$\sum_{j=1}^{p} t_j(\boldsymbol{B}'\boldsymbol{X})=\mathrm{tr}\,\boldsymbol{\Sigma}-\mathrm{tr}\,\boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{\Sigma}.$$

欲证(7.74)，只需证

$$\mathrm{tr}\,\boldsymbol{\Sigma}\boldsymbol{U}_1(\boldsymbol{U}_1'\boldsymbol{\Sigma}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1'\boldsymbol{\Sigma}=\max_{\mathrm{rk}\boldsymbol{B}=k}\mathrm{tr}\,\boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{\Sigma} \tag{7.75}$$

注意到

$$\mathrm{tr}\,\boldsymbol{\Sigma}\boldsymbol{U}_1(\boldsymbol{U}_1'\boldsymbol{\Sigma}\boldsymbol{U}_1)^{-1}\boldsymbol{U}_1'\boldsymbol{\Sigma}=\mathrm{tr}\,\boldsymbol{U}_1\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\Lambda}_k\boldsymbol{U}_1'=\mathrm{tr}\,\boldsymbol{U}_1\boldsymbol{\Lambda}_k\boldsymbol{U}_1'=\sum_{1}^{k}\lambda_i.$$

其中 $\Lambda_k=\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_k\end{pmatrix}$, λ_i 是 $\boldsymbol{\Sigma}$ 的第 i 个顺序特征值. 由于 $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{\Sigma}^{\frac{1}{2}}$ 是到 $\mu(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B})$ 的正投影阵，由 A.6.2，存在正交阵 V 的前 k 列 V_1 满足

$$\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{\Sigma}^{\frac{1}{2}}=\boldsymbol{V}_1\boldsymbol{V}_1',\quad \boldsymbol{V}_1'\boldsymbol{V}_1=\boldsymbol{I}_k,$$

于是 $\operatorname{tr}\boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{\Sigma}=\operatorname{tr}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{V}_1\boldsymbol{V}_1'\boldsymbol{\Sigma}^{\frac{1}{2}}=\operatorname{tr}\boldsymbol{V}_1'\boldsymbol{\Sigma}\boldsymbol{V}_1$,

根据 A.4.3 得 $\max\limits_{V_1'V_1=I_k}\operatorname{tr}\boldsymbol{V}_1'\boldsymbol{\Sigma}\boldsymbol{V}_1=\sum\limits_{i=1}^{k}\lambda_i$ 在 $\boldsymbol{V}_1=\boldsymbol{U}_1$ 时达到. 证毕.

记 $\widetilde{\boldsymbol{x}}=\boldsymbol{U}_1\widetilde{\boldsymbol{y}}$, 这里 $\widetilde{\boldsymbol{y}}=(y_1,\ \cdots,\ y_k)'$. 我们有

$$\widetilde{X}_i=\sum_{j=1}^{k}u_{ij}y_j,\ D(\widetilde{X}_i)=\sum_{j=1}^{k}u_{ij}^2\lambda_j,$$

而

$$X_i=\sum_{j=1}^{p}u_{ij}y_j,\ D(X_i)=\sum_{j=1}^{p}u_{ij}^2\lambda_j,$$

从而可见 $\widetilde{X}_i$ 的方差是 X_i 的方差的一部分, 其所占比例为 $\sum\limits_{j=1}^{k}u_{ij}^2\lambda_j\Big/\sum\limits_{j=1}^{p}u_{ij}^2\lambda_j$, 这个比例如果较接近于 1, 就意味着用前 k 个主成分的信息, 能基本上"包含"X_i 的信息. 由于在比例式中起作用的不仅仅是 λ_j, u_{ij}^2 亦可起显著影响, 故称 u_{ij} 为 X_i 在主成分 Y_j 上的**载荷**. 对于载荷的分析, 在实际应用时是很重要的.

应用中常考虑将**主成分标准化**, 即将 y_i 除以它的标准差 $\sqrt{\lambda_i}$, 从而使 $D(Y_i/\sqrt{\lambda_i})=1$. 记 $f_i=Y_i/\sqrt{\lambda_i}$, 则有 $\operatorname{Cov}(\boldsymbol{f})=\boldsymbol{I}_k$. 由

$$\widetilde{\boldsymbol{X}}=\boldsymbol{U}_1\begin{pmatrix}\sqrt{\lambda_1} & & 0\\ & \ddots & \\ 0 & & \sqrt{\lambda_k}\end{pmatrix}\boldsymbol{f}=\boldsymbol{U}_1\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{f}$$

可得 X_i 在标准化主成分 f_j 上的载荷为 $u_{ij}\sqrt{\lambda_j}$, 记 $a_{ij}=u_{ij}\sqrt{\lambda_j}$. 有 $D(\widetilde{X}_i)=\sum\limits_{j=1}^{k}a_{ij}^2$, $D(X_i)=\sum\limits_{j=1}^{p}a_{ij}^2$, 故对标准化主成分而言, 载荷的意义更为明显. 事实上我们有: 在 $\boldsymbol{\Sigma}=\boldsymbol{R}$ 是相关矩阵时,

$$a_{ij}=\rho(X_i,\ Y_j)\quad (X_i \text{ 与 } Y_j \text{ 的相关系数}).$$

还可以考虑将 $\boldsymbol{f}$ 旋转, 即以 k 阶正交阵 $\boldsymbol{\Gamma}$ 去作用, 令 $\boldsymbol{g}=\boldsymbol{\Gamma}'\boldsymbol{f}$, 则有 $\widetilde{\boldsymbol{X}}=\boldsymbol{U}_1\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Gamma}\boldsymbol{g}$, 仍有 $D(\widetilde{X}_i)=\sum\limits_{j=1}^{k}a_{ij}^2$, 但这时 X_i 在 g_j 上的载荷变为

$$b_{ij}\triangleq\sum_{t=1}^{k}a_{it}\gamma_{tj},\ \text{其中}\ \gamma_{tj}\ \text{是}\ \Gamma\ \text{的}(t,\ j)\text{元}.$$

适当选取 Γ, 有可能对问题作出更好的统计解释.

在实际进行主成分分析时，总体的协方差阵或相关矩阵是不大可能知道的，那时总是从 n 个 p 维观察值出发，用样本协方差阵 $\hat{\Sigma}$ 或样本相关矩阵 $\hat{R}$ 去进行分析，在寻求主成分的程序上并无差别，但在对结果作理论分析时涉及到的特征值、特征向量都是随机变量，这就使推导变得十分困难．特征值的分布虽可在一定条件下导出，但已超出本书的范围．

在使用样本协方差阵作主成分分析时，对主成分可作出有趣的几何解释．设 n 个观察点为 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$，我们取这 n 个点的重心 $\frac{1}{n}\sum_1^n \boldsymbol{x}_i$ 为原点，即假定数据阵 $\boldsymbol{X} \triangleq \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}$ 的各列都以该列的平均值为起始点，即以 $x_{ij}-\frac{1}{n}\sum_{k=1}^n x_{ik}$ 代替 x_{ij}．这时样本协方差阵就是 $\frac{1}{n-1}\boldsymbol{X}'\boldsymbol{X}$，记 $\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{C}$，现在我们要找一条直线（通过原点，方向为 u，$\boldsymbol{u}'\boldsymbol{u}=1$），使 n 个点 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ 到此直线的距离平方和最小．容易看出 $\boldsymbol{x}_i$ 到此直线的距离平方是 $\boldsymbol{x}_i'\boldsymbol{x}_i-\boldsymbol{x}_i'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{x}_i$．因此，问题就是要极小化

$$S_n(\boldsymbol{X},\ \boldsymbol{u}) \triangleq \sum_{i=1}^n(\boldsymbol{x}_i'\boldsymbol{x}_i-\boldsymbol{x}_i'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{x}_i)=\operatorname{tr}\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{u}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{u}.$$

上式的极小值点显然是 $\boldsymbol{X}'\boldsymbol{X}$ 的相应于最大特征值的特征向量 $\boldsymbol{u}_1$．故由 $\boldsymbol{u}_1'\boldsymbol{x}$ 为第一主成分的直观理由就很明显，即数据在这个方向上有最大的变异．接着就是在与 $\boldsymbol{u}_1$ 正交的线性空间（当原点不是重心时，是通过重心与 $\boldsymbol{u}_1$ 正交的超平面）中以类似上述方法找一条直线，这条直线的方向 $\boldsymbol{u}_2(\boldsymbol{u}_2'\boldsymbol{u}_2=1)$ 正好对应着第二主成分 $\boldsymbol{u}_2'\boldsymbol{X}$，如此等等．

顺便指出，这里问题的提法与线性回归不同．如考虑 $\boldsymbol{X}_1$ 对 $\boldsymbol{X}_2, \cdots, \boldsymbol{X}_p$ 的线性回归，它要寻求一个超平面（$p-1$ 维）使得 n 个点沿着第一个坐标轴的方向到此超平面的距离平方和最小．它是沿一个事先固定的方向计算距离，与主成分分析的几何解释中计算点到超平面的垂直距离不同．这在 $p=2$ 时很容易用图形说明．

(二)因子分析

因子分析方法被认为开端于1904年C. Spearman的著名论文. 他在分析中学一个班级的学生的六门功课成绩时发现，六个随机变量实际上可分解成如下形式

$$\boldsymbol{x} \triangleq \begin{pmatrix} x_1 \\ \vdots \\ x_6 \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_6 \end{pmatrix} f + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_6 \end{pmatrix}. \tag{7.76}$$

这里f是对每个x_i都起作用的变量，ε_i只对x_i起作用，a_i是常数，称这样的f为$\boldsymbol{x}$的**公共因子**，ε_i是**特殊因子**. 又如考虑人的生理指标. 设为收缩压、舒张压、心跳间隔、呼吸间隔、舌下温度等五项，从生理学获知这些指标至少受植物神经的影响，而植物神经又分交感神经和副交感神经，因此这两项可作为以上生理指标的公共因子，至于还有没有其它公共因子，则需作进一步分析. 一般地有

$$\underset{p\times 1}{\boldsymbol{x}} = \underset{p\times q\times 1}{\boldsymbol{A}\boldsymbol{f}} + \boldsymbol{\varepsilon},\ f\text{——公共因子},\ \varepsilon\text{——特殊因子}. \tag{7.77}$$

并且有理由假定:

(i) $q \leqslant p$;

(ii) $\boldsymbol{f}$与$\boldsymbol{\varepsilon}$不相关，即

$$\mathrm{Cov}(\boldsymbol{f},\ \boldsymbol{\varepsilon}) = 0; \tag{7.78}$$

(iii) $\mathrm{Cov}\boldsymbol{f} = \boldsymbol{I}_q$, $\mathrm{Cov}\boldsymbol{\varepsilon} = \mathrm{diag}(\sigma_1^2,\ \cdots,\ \sigma_p^2) \triangleq \boldsymbol{\Delta}$.

这类假定主要出于数学处理的需要，但在不少情形下也可由实际问题作出解释.

对指标进行这样的因子分析，从而明确对所有指标都有影响的某些公共因子，并确定只对某个指标起作用的特殊因子，在实践中当然是有意义的，这使得我们在处理有关问题时能采取更合理的措施，并常可使维数降低.

下面我们对满足假定(7.78)的模型(7.77)作一些理论探讨.

计算$\boldsymbol{x}$的协方差阵，有

$$\mathrm{Cov}\,\boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}' + \boldsymbol{\Delta}. \tag{7.79}$$

可见

$$D(x_i)=\sum_{j=1}^{q}a_{ij}^2+\sigma_i^2,$$

得 x_i 的方差是由两部分组成的. 一部分是 A 的第 i 行向量 $\boldsymbol{a}(i)$ 的范数平方 $\|a_{(i)}\|^2$, 它由公共因子的系数矩阵决定, 另一部分 σ_i^2 是第 i 个特殊因子的方差. 如果 σ_i^2 较小, 则 $D(x_i)$ 主要由公共因子的影响决定. 记 $h_i^2=\|\boldsymbol{a}_{(i)}\|^2$, 用它来表明公共因子对 x_i 的影响的大小. 在极端情形, 当 $\sigma_i^2=0$, 表明 x_i 完全由公共因子决定; 当 $h_i^2=0$, A 的第 i 行为 0, x_i 是不依赖公共因子的, 它完全由特殊因子决定. 称 h_i^2 **为公共因子 $\boldsymbol{f}$ 对 x_i 的贡献**. 考虑公共因子中 f_j 对 x 的影响不难看出, 这种影响是通过 A 的第 j 列表现的, 记 $g_j^2=\sum_{i=1}^{p}a_{ij}^2$, 称 g_j^2 **为公共因子中的 f_j 对 $\boldsymbol{x}$ 的贡献**. 如同主成分分析中相仿, 称 a_{ij} 为第 i 个指标 x_i 在第 j 个公共因子 f_j 上的**载荷**. 使 g_j^2 达最大的公共因子 f_j 是最重要的公共因子, 使 h_i^2 达最大的指标 x_i 是对公共因子依赖最大的指标. 而载荷 a_{ij}, 在 $D(x_i)=D(f_j)=1$ 时, 恰好是 x_i 与 f_j 的相关系数 $\rho(x_i, f_j)$. 在实际工作中往往希望载荷相对集中, 这样便于分析指标和公共因子的关系. 习惯上称 $\boldsymbol{A}$ 为**载荷矩阵**. 为方便起见, 我们不妨设它的列已按对 $\boldsymbol{x}$ 的贡献的大小次序排好, 即 $g_1^2\geqslant\cdots\geqslant g_q^2$.

一个关键的理论问题是求出载荷矩阵 $\boldsymbol{A}$. 在较强的假定下, 这是容易做到的. 设 $\boldsymbol{x}$ 已经标准化(即使每个分量的方差为 1), 它的协方差矩阵 $\mathrm{Cov}\,\boldsymbol{x}=\boldsymbol{R}$ 就是相关矩阵. 设 $\boldsymbol{R}$ 和 $\boldsymbol{\Delta}$ 已知, 则有 $\boldsymbol{R}_*\triangleq\boldsymbol{R}-\boldsymbol{\Delta}$ 称为**约相关阵**. 由约相关阵 $\boldsymbol{R}_*=\boldsymbol{A}\boldsymbol{A}'$ 不难定出 $\boldsymbol{A}$ 满足前面的要求. 在这种情形下我们当然要有 $\boldsymbol{R}_*\geqslant0$(非负定), 用附录 A.4.3 中的谱分解, 有

$$\boldsymbol{R}_*=\sum_{j=1}^{r}\lambda_j\boldsymbol{u}_j\boldsymbol{u}_j',\ r=\mathrm{rk}\,\boldsymbol{R}_*,$$

其中 $\boldsymbol{u}_j$ 是 $\boldsymbol{R}_*$ 的相应于特征值 λ_j 的规范化特征向量(即 $\|\boldsymbol{u}_j\|=1$), $\lambda_1\geqslant\cdots\geqslant\lambda_r>0$. 于是, 取

$$\boldsymbol{a}_j=\sqrt{\lambda_j}\,\boldsymbol{u}_j,\ j=1,\ \cdots,\ q, \text{就有 } \boldsymbol{A}=(\boldsymbol{a}_1,\ \cdots\boldsymbol{a}_q)$$

为欲求的系数矩阵. 这样求得的 $\boldsymbol{A}$ 还满足

$$\boldsymbol{A}'A=\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_q\end{pmatrix}, \text{有 } g_j^2=\lambda_j,\ j=1,\ \cdots,\ q.$$

在上述情形下，因子分析似与主成分分析无多大差别. 但在主成分分析中，是去寻求相关矩阵 $\boldsymbol{R}$ 的前 k 个特征向量，而在因子分析中则从约相关阵 $\boldsymbol{R}_*$ 出发，这是不同的. 为了保证 $\boldsymbol{R}_*$ 是非负定阵，我们必须限制 $\boldsymbol{\Delta}=\mathrm{diag}(\sigma_1^2,\ \cdots,\ \sigma_p^2)$在一定的范围内. 并且，在主成分分析中要寻求的 $\boldsymbol{U}_1$ 是各列正交的，但对 $\boldsymbol{A}$ 却无此约束，因而这里的解是不完全确定的.

根据 $\boldsymbol{R}_*=\boldsymbol{A}\boldsymbol{A}'$ 去求 $\boldsymbol{A}$，解虽然并不唯一，但由附录 A.4.6，如 $\boldsymbol{B}\boldsymbol{B}'=\boldsymbol{A}\boldsymbol{A}'$，则有 $\boldsymbol{B}=\boldsymbol{A}\boldsymbol{\Gamma}$，这里 $\boldsymbol{\Gamma}$ 是 q 维正交阵. 因此解的空间 $\mu(\boldsymbol{A})$是唯一的，称为因子空间，而且任一解均可由上面已求得的 A 经过旋转(右乘一正交阵)而得. 经过旋转之后，公共因子对 x_i 的贡献 h_i^2 并不改变，但公共因子本身就可能有较大变化，即 g_j^2 不再与原来相同，从而可通过适当的旋转去得到使我们比较称心的公共因子.

经过旋转，可将模型(7.77)记为

$$\boldsymbol{x}=(\boldsymbol{A}\boldsymbol{\Gamma})(\boldsymbol{\Gamma}'\boldsymbol{f})+\varepsilon\triangleq\boldsymbol{B}\boldsymbol{y}+\varepsilon \tag{7.80}$$

这里 $\boldsymbol{y}=\boldsymbol{\Gamma}'\boldsymbol{f}$.

如同在主成分分析中相仿，我们希望各个因子的贡献越"分散"越好，即尽可使得有较多的载荷为零或接近于零，使各个因子的影响分别体现在部分指标上. 注意到 $\sum_1^q g_j^2=\mathrm{tr}\,\boldsymbol{A}\boldsymbol{A}'$ 是不因旋转而影响的，因此，旋转后贡献的分散程度可由各个列的样本方差来体现. 实际做法以 $q=2$ 为例: 设

$$\boldsymbol{B}=\boldsymbol{A}\boldsymbol{\Gamma}=\begin{pmatrix}b_{11} & b_{12}\\ \vdots & \\ b_{p1} & b_{p2}\end{pmatrix}. \tag{7.81}$$

为了消除符号不同的影响，并消除各个指标对公共因子的依赖程度的影响，考虑以 b_{ij}^2/h_i^2 代替 b_{ij}，令

$$S_j=\frac{1}{p}\sum_{i=1}^{p}\left[\left(\frac{b_{ij}^2}{h_i^2}\right)-\frac{1}{p}\sum_{i=1}^{p}\frac{b_{ij}^2}{h_i^2}\right]^2,\ j=1,\ 2, \tag{7.82}$$
$$S=S_1+S_2.$$

寻求 $\boldsymbol{\Gamma}$ 使 S 极大化. 这样的旋转称为**方差最大的正交旋转**. 因为 S 仅仅是旋转角的弧度 φ 的函数, 这样的正交旋转不难求出, 为节省篇幅留作练习. 当 $q>2$ 时, 可配对逐步进行, 必要时反复若干次, 尽可能取得较满意的效果. 这方面的实例可参看肯德尔《多元分析》第 4 章.

(三) 典型相关

前面我们讨论了随机向量分量的互依性, 给出了主成分分析和因子分析方法. 在本段中, 我们将讨论两个随机向量之间的互依性.

在刻划两个随机变量的互依性时, 引进了相关系数的概念, 它反映两个变量线性相关的程度. 在讨论回归时, 我们引进了变量 Y 和向量 X 之间线性相依程度的刻划量, 称为多重相关系数或称复相关系数. Y 和 X 的复相关系数定义为 Y 和 $\boldsymbol{X}$ 的线性函数 $\boldsymbol{a}'\boldsymbol{X}$ 间的相关系数 $\rho(Y,\ \boldsymbol{a}'\boldsymbol{X})$ 的极大值, 记为 $\rho_{Y\cdot\boldsymbol{X}}$. 当 $\rho_{Y\cdot\boldsymbol{X}}$ 较大时, 用 Y 对 $\boldsymbol{X}$ 的线性回归来预测 Y 的精度就高; 反之, Y 和 $\boldsymbol{X}$ 的线性相依性就不紧密, Y 就不宜用 $\boldsymbol{X}$ 的线性函数来预测. 可见线性相依性的研究有实际意义.

把线性相依性推广到两个随机向量之间, 应归功于 Hotelling (1936). 他引进了典型相关的概念如下:

定义 7.4

设 $\boldsymbol{X}=(X_1,\ \cdots,\ X_p)'$ 和 $\boldsymbol{Y}=(Y_1,\ \cdots,\ Y_q)'$ 是两个随机向量, $\boldsymbol{a}'\boldsymbol{X}$ 和 $\boldsymbol{b}'\boldsymbol{Y}$ 是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的两个任意的线性函数, 为确定起见, 设它们的方差都是 1, 即 $D(\boldsymbol{a}'\boldsymbol{X})=D(\boldsymbol{b}'\boldsymbol{Y})=1$. 记 $\rho(\boldsymbol{a}'\boldsymbol{X},\ \boldsymbol{b}'\boldsymbol{Y})$ 是 $\boldsymbol{a}'\boldsymbol{X}$ 与 $\boldsymbol{b}'\boldsymbol{Y}$ 的相关系数, 如果 $\boldsymbol{a}_1',\ \boldsymbol{b}_1'$ 满足 $D(\boldsymbol{a}_1'\boldsymbol{X})=D(\boldsymbol{b}_1'\boldsymbol{Y})=1$, 且

$$\rho(\boldsymbol{a}_1'\boldsymbol{X},\ \boldsymbol{b}_1'\boldsymbol{Y})=\max_{D(\boldsymbol{a}'\boldsymbol{X})=D(\boldsymbol{b}'\boldsymbol{Y})=1}\rho(\boldsymbol{a}'\boldsymbol{X},\ \boldsymbol{b}'\boldsymbol{Y}) \tag{7.83}$$

则称 $\rho(\boldsymbol{a}_1'\boldsymbol{X},\ \boldsymbol{b}_1'\boldsymbol{Y})$是 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的(第一)典型相关系数,称 $\boldsymbol{a}_1'\boldsymbol{X}$ 和 $\boldsymbol{b}_1'\boldsymbol{Y}$ 是 $\boldsymbol{X}$, $\boldsymbol{Y}$ 的(第一组)典型(相关)变量,在不致混淆时简记 $\rho_1=\rho(\boldsymbol{a}_1'\boldsymbol{X},\ \boldsymbol{b}_1'\boldsymbol{Y})$.

从实际观点看,ρ_1 反映了基于 $\boldsymbol{X}$ 的综合(线性组合)指标和基于 $\boldsymbol{Y}$ 的综合指标间的最大相关程度,它也是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的互依性的刻划.(这里的互依性是指某种没有精确定义的线性互依性,因为我们并未讨论 $\boldsymbol{X}$ 的任意函数和 $\boldsymbol{Y}$ 的任意函数之间的相关系数.)

在 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的联合二阶矩已知时,典型相关系数和典型相关变量组的推导是不难的.事实上,设

$$\mathrm{Cov}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY}\\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY}\end{pmatrix},\ \boldsymbol{\Sigma}_{XX},\ \boldsymbol{\Sigma}_{YY}>0 \tag{7.84}$$

任给线性函数 $\boldsymbol{a}'\boldsymbol{X}$ 与 $\boldsymbol{b}'\boldsymbol{Y}$,有

$$\rho(\boldsymbol{a}'\boldsymbol{X},\ \boldsymbol{b}'\boldsymbol{Y})=\boldsymbol{a}'\boldsymbol{\Sigma}_{XY}\boldsymbol{b}=\boldsymbol{b}'\boldsymbol{\Sigma}_{YX}\boldsymbol{a}. \tag{7.85}$$

它在条件 $\boldsymbol{a}'\boldsymbol{\Sigma}_{XX}\boldsymbol{a}=\boldsymbol{b}'\boldsymbol{\Sigma}_{YY}\boldsymbol{b}=1$(即 $D(\boldsymbol{a}'\boldsymbol{X})=D(\boldsymbol{b}'\boldsymbol{Y})=1$)下的极大值显然可用 Lagrange λ-乘子法求出.令

$$\varphi(\boldsymbol{a},\ \boldsymbol{b})=\boldsymbol{a}'\boldsymbol{\Sigma}_{XY}\boldsymbol{b}-\frac{\lambda}{2}(\boldsymbol{a}'\boldsymbol{\Sigma}_{XX}\boldsymbol{a}-1)-\frac{\mu}{2}(\boldsymbol{b}'\boldsymbol{\Sigma}_{YY}\boldsymbol{b}-1).$$

容易验算

$$\frac{\partial\varphi}{\partial\boldsymbol{a}}\triangleq\begin{pmatrix}\dfrac{\partial\varphi}{\partial a_1}\\ \vdots\\ \dfrac{\partial\varphi}{\partial a_p}\end{pmatrix}=\boldsymbol{\Sigma}_{XY}\boldsymbol{b}-\lambda\boldsymbol{\Sigma}_{XX}\boldsymbol{a},$$

$$\frac{\partial\varphi}{\partial\boldsymbol{b}}\triangleq\begin{pmatrix}\dfrac{\partial\varphi}{\partial b_1}\\ \vdots\\ \dfrac{\partial\varphi}{\partial b_q}\end{pmatrix}=\boldsymbol{\Sigma}_{YX}\boldsymbol{a}-\mu\boldsymbol{\Sigma}_{YY}\boldsymbol{b}.$$

令 $\dfrac{\partial\varphi}{\partial\boldsymbol{a}}=\boldsymbol{0}$, $\dfrac{\partial\varphi}{\partial\boldsymbol{b}}=\boldsymbol{0}$, 在原条件下易见

$$\lambda=\mu=\boldsymbol{a}'\boldsymbol{\Sigma}_{XY}\boldsymbol{b}\triangleq\rho,$$

并得

$$\begin{aligned} \boldsymbol{W}_1\boldsymbol{a} &\triangleq \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{a}=\rho^2\boldsymbol{a}, \\ \boldsymbol{W}_2\boldsymbol{a} &\triangleq \boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{b}=\rho^2\boldsymbol{b}. \end{aligned} \tag{7.86}$$

由附录 A.5.4 知 $\boldsymbol{W}_1$ 与 $\boldsymbol{W}_2$ 有相同的非零特征值. (7.86)表明 ρ^2 是 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 的特征值，而 $\boldsymbol{a}$, $\boldsymbol{b}$ 分别是 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 的相应于特征值 ρ^2 的特征向量. 设 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 的非零特征值的个数(包括重数)是 r 个. 则 $\varphi(a, b)$有 r 个稳定点，有 r 个局部极值 $|\rho_1|\geqslant\cdots\geqslant|\rho_r|>0$，其中 $|\rho_1|$ 为极大值. 它正是我们欲求的 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的典型相关系数，而相应于 ρ_1^2 的 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 的特征向量 $\boldsymbol{a}_1$, $\boldsymbol{b}_1$ 就给出了典型相关变量组 $\boldsymbol{a}_1'\boldsymbol{X}$, $\boldsymbol{b}_1'\boldsymbol{Y}$(注意我们约定了 $D(\boldsymbol{a}_1'\boldsymbol{X})=D(\boldsymbol{b}_1'\boldsymbol{Y})=1$). 这里 $\boldsymbol{a}_1$, $\boldsymbol{b}_1$ 除方向可以相反外一般是确定的(当 $\rho_1^2=\rho_2^2$，它们就可在一个 2 维的线性空间中任取，但这种情形出现的概率为零，可予忽略). 按定义典型相关系数应取正值.

在应用时，一组典型相关变量可能不足以反映 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的互依性，有时需要讨论它们的较多个综合指标，为此，我们称 $|\rho_i|$ 为第 i 个典型相关系数，相应于 ρ_i 的 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 的特征向量 $\boldsymbol{a}_i$, $\boldsymbol{b}_i$ 给出的变量 $\boldsymbol{a}_i'\boldsymbol{X}$, $\boldsymbol{b}_i'\boldsymbol{Y}$ 称作第 i 组典型相关变量，$i=1,\cdots,r$. 为方便计，取 ρ_i 为正值.

可在 $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_r$ 后添加 $\boldsymbol{a}_{r+1}, \cdots, \boldsymbol{a}_p$，使得 $\{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{a}_1,\cdots,\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{a}_p\}$ 是 $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{W}_1\boldsymbol{\Sigma}^{-\frac{1}{2}}$ 的正交规范化特征向量集，对 $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_r$ 亦添加 $\boldsymbol{b}_{r+1},\cdots, \boldsymbol{b}_q$，使有上述类似性质. 记

$$\boldsymbol{A}=(\boldsymbol{a}_1\cdots\boldsymbol{a}_p),\ \boldsymbol{B}=(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_q), \tag{7.87}$$

注意到上述取法有

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{A}'\boldsymbol{X}) &= \boldsymbol{A}'\boldsymbol{\Sigma}_{XX}\boldsymbol{A}=\boldsymbol{I}_p, \\ \mathrm{Cov}(\boldsymbol{B}'\boldsymbol{X}) &= \boldsymbol{B}'\boldsymbol{\Sigma}_{YY}\boldsymbol{B}=\boldsymbol{I}_q. \end{aligned}$$

由于 $\boldsymbol{a}_i'\boldsymbol{\Sigma}_{XY}\boldsymbol{b}_j=\rho_j\boldsymbol{a}_i'\boldsymbol{\Sigma}_{XX}\boldsymbol{a}_i=0$，又得

$\mathrm{Cov}(\boldsymbol{A}'\boldsymbol{X}, \boldsymbol{B}'\boldsymbol{Y})=\boldsymbol{\Delta}$，除主对角元 $\rho_1, \cdots\ \rho_r$ 外，其余皆为 0. 从而有

$$\mathrm{Cov}\begin{pmatrix} \boldsymbol{A}'\boldsymbol{X} \\ \boldsymbol{B}'\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_p & \boldsymbol{\Delta} \\ \boldsymbol{\Delta}' & \boldsymbol{I}_q \end{pmatrix} \tag{7.88}$$

因此，寻求典型变量实际上就是对原变量作线性变换，使变换后的变量组具有如(7.88)的简单协方差结构．故在研究互依性时，我们至少可以抛弃那些使相应的新变量组协方差为零的部分，从而也起了使互依性突出表现于若干组典型变量的作用，这也就降低了维数．

在实际应用典型相关时，我们可将相对很小的 ρ_i 所对应的典型变量抛弃，而只考虑 ρ_i 较大的若干组典型变量，设为前 k 组．记 $\boldsymbol{A}_1$, $\boldsymbol{B}_1$ 分别是 $\boldsymbol{A}$, $\boldsymbol{B}$ 的前 k 列．我们可以用 $\boldsymbol{A}_1'\boldsymbol{X}$ 和 $\boldsymbol{B}_1'\boldsymbol{Y}$ 的相关来近似地反映 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的相关．

典型变量的另一统计解释是：若要用 $\boldsymbol{Y}$ 的线性函数来预测 $\boldsymbol{a}_i'\boldsymbol{X}$, $\boldsymbol{a}_i$ 如(7.87)，则使得均方误差最小的线性预测是

$$\widehat{\boldsymbol{a}_i'\boldsymbol{X}} = \boldsymbol{a}_i'E\boldsymbol{X} - \rho_i\boldsymbol{b}_i'E\boldsymbol{Y} + \rho_i\boldsymbol{b}_i'\boldsymbol{Y}. \tag{7.89}$$

事实上，设 $C_0 + \boldsymbol{C}'\boldsymbol{Y}$ 是欲求的预测，由(6.18)得 $\widehat{\boldsymbol{a}_i'\boldsymbol{X}} = \boldsymbol{a}_i'E\boldsymbol{X} - \boldsymbol{a}_i'\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}E\boldsymbol{Y} + \boldsymbol{a}_i'\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{Y}$, 由于 $\boldsymbol{\Sigma}_{YX}\boldsymbol{a}_i = \rho_i\boldsymbol{\Sigma}_{YY}\boldsymbol{b}_i$, 得(7.89)．

典型变量的一个有意思的应用是用来给出 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的公共因子 $\boldsymbol{Z}$．即存在 $\boldsymbol{Z}$ 满足

$$\begin{cases} \boldsymbol{X} = \boldsymbol{C}_1\boldsymbol{Z} + \boldsymbol{\varepsilon}_1, \\ \boldsymbol{Y} = \boldsymbol{C}_2\boldsymbol{Z} + \boldsymbol{\varepsilon}_2. \end{cases} \tag{7.90}$$

且有　$\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{\varepsilon}_1) = 0$, $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{\varepsilon}_2) = 0$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2) = 0$.

事实上，(7.90)中的 $\boldsymbol{Z}$ 可取典型变量 $\boldsymbol{A}_1'\boldsymbol{X}$, $\boldsymbol{A}_1$ 是(7.87)中 $\boldsymbol{A}$ 的前 r 列，$\boldsymbol{B}_1$ 是 $\boldsymbol{B}$ 的前 r 列．注意到 $\boldsymbol{A}'\boldsymbol{\Sigma}_{XX}\boldsymbol{A} = \boldsymbol{I}_p$, $\boldsymbol{B}'\boldsymbol{\Sigma}_{YY}\boldsymbol{B} = \boldsymbol{I}_q$, 有

$$\boldsymbol{\Sigma}_{XX}\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}_p, \quad \boldsymbol{\Sigma}_{YY}\boldsymbol{B}\boldsymbol{B}' = \boldsymbol{I}_q.$$

故若记 $\boldsymbol{\Lambda} = \mathrm{Cov}(\boldsymbol{A}_1'\boldsymbol{X}, \boldsymbol{B}_1'\boldsymbol{Y})$，就有

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{\Sigma}_{XX}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{X} \\ &= \boldsymbol{\Sigma}_{XX}\boldsymbol{A}_1\boldsymbol{A}_1'\boldsymbol{X} + \boldsymbol{\Sigma}_{XX}\boldsymbol{A}_2\boldsymbol{A}_2'\boldsymbol{X}; \end{aligned}$$

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{\Sigma}_{YY}\boldsymbol{B}_1\boldsymbol{B}_1'\boldsymbol{Y} + \boldsymbol{\Sigma}_{YY}\boldsymbol{B}_2\boldsymbol{B}_2'\boldsymbol{Y} \\ &= \boldsymbol{\Sigma}_{YY}\boldsymbol{B}_1\boldsymbol{B}_1'\boldsymbol{\Lambda}\boldsymbol{A}_1'\boldsymbol{X} + \boldsymbol{\Sigma}_{YY}\boldsymbol{B}_1\boldsymbol{B}_1'(\boldsymbol{Y} - \boldsymbol{\Lambda}\boldsymbol{A}_1'\boldsymbol{X}) \\ &\quad + \boldsymbol{\Sigma}_{YY}\boldsymbol{B}_2\boldsymbol{B}_2'\boldsymbol{Y}, \end{aligned}$$

记 $\boldsymbol{Z}=\boldsymbol{A}_1'\boldsymbol{X}$, $\boldsymbol{C}_1=\boldsymbol{\Sigma}_{XX}\boldsymbol{A}_1$, $\boldsymbol{C}_2=\boldsymbol{\Sigma}_{YY}\boldsymbol{B}_1\boldsymbol{B}_1'\boldsymbol{\Lambda}$, $\boldsymbol{\varepsilon}_1=\boldsymbol{\Sigma}_{XX}\boldsymbol{A}_2\boldsymbol{A}_2'\boldsymbol{X}$, $\boldsymbol{\varepsilon}_2=\boldsymbol{\Sigma}_{YY}\boldsymbol{B}_1\boldsymbol{B}_1'(\boldsymbol{Y}-\boldsymbol{\Lambda}\boldsymbol{A}_1'\boldsymbol{X})+\boldsymbol{\Sigma}_{YY}\boldsymbol{B}_2\boldsymbol{B}_2'\boldsymbol{Y}$, 就有 $\boldsymbol{X}$, $\boldsymbol{Y}$ 如(7.90), 并且容易验算

$$\mathrm{Cov}(\boldsymbol{A}_1'\boldsymbol{X},\ \boldsymbol{\Sigma}_{XX}\boldsymbol{A}_2\boldsymbol{A}_2'\boldsymbol{X})=\boldsymbol{A}_1'\boldsymbol{\Sigma}\boldsymbol{A}_2\boldsymbol{A}_2'\boldsymbol{\Sigma}_{XX}=\boldsymbol{0},$$

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{A}_1'\boldsymbol{X},\ \boldsymbol{\varepsilon}_2)&=\boldsymbol{A}_1'\boldsymbol{\Sigma}_{XY}\boldsymbol{B}_1\boldsymbol{B}_1'\boldsymbol{\Sigma}_{YY}-\boldsymbol{A}_1'\boldsymbol{\Sigma}_{XX}\boldsymbol{A}_1\boldsymbol{\Lambda}\boldsymbol{B}_1'\boldsymbol{\Sigma}_{YY}+\boldsymbol{0}\\&=\boldsymbol{\Lambda}\boldsymbol{B}_1'\boldsymbol{\Sigma}_{YY}-\boldsymbol{\Lambda}\boldsymbol{B}_1'\boldsymbol{\Sigma}_{YY}=\boldsymbol{0},\end{aligned}$$

$$\mathrm{Cov}(\boldsymbol{\varepsilon}_1,\ \boldsymbol{\varepsilon}_2)=\mathrm{Cov}(\boldsymbol{\Sigma}_{XX}\boldsymbol{A}_2\boldsymbol{A}_2'\boldsymbol{X},\ \boldsymbol{Y}-\boldsymbol{\Sigma}_{YY}\boldsymbol{B}_1\boldsymbol{B}_1'\boldsymbol{\Lambda}\boldsymbol{A}_1\boldsymbol{X})=\boldsymbol{0}.$$

故得 $\boldsymbol{Z}$ 是 $\boldsymbol{X}$, $\boldsymbol{Y}$ 的公共因子.

顺便指出,在主成分分析和因子分析中用过的旋转因子法,在典型相关分析中已不适用. 原因是这里对 $\boldsymbol{X}$, $\boldsymbol{Y}$ 的变换要求联合协方差阵有结构(7.88), 但任何旋转都将使协方差阵失去那种简单结构.

从样本观察值出发的因子分析和典型相关分析, 可仿照(一)中所述的精神进行,不在此一一讨论.

习　　题

1. 考虑下列样本($n=8$)

谷物重量(y_1)	40	17	9	15	6	12	5	9
稻草重量(y_2)	53	19	10	29	13	27	19	30
施 肥 量	24	11	5	12	7	14	11	18

设 $x_{1\alpha}=1$, $x_{2\alpha}$ 是在第 α 个区组的施肥量, $\alpha=1, \cdots, 8$. 估计 $\boldsymbol{B}=\begin{pmatrix}\boldsymbol{B}_1\\\boldsymbol{B}_2\end{pmatrix}$, 在 0.01 显著性水平上检验 $\boldsymbol{B}_2=0$.

2. 证明 § 7.1(一)中性质 5.1°.

3. 证明定理 7.1.

4. 设 $\boldsymbol{x}\sim N_{np}(\boldsymbol{M},\ \boldsymbol{I},\ \boldsymbol{\Sigma})$, $\boldsymbol{A}$, $\boldsymbol{B}$ 是 n 阶对称阵, 且 $\boldsymbol{AB}=\boldsymbol{0}$, 证明: $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ 与 $\boldsymbol{y}'\boldsymbol{B}\boldsymbol{y}$ 独立.

5. 试证: 当 $\boldsymbol{x}\sim N_{np}(\boldsymbol{M},\ \boldsymbol{I},\ \boldsymbol{\Sigma})$, $\boldsymbol{x}'\boldsymbol{x}$ 的分布只通过 $\boldsymbol{M}'\boldsymbol{M}$ 依赖 $\boldsymbol{M}$.

6. 将 Cochran 定理推广到 $\boldsymbol{x}\sim N_{np}(\boldsymbol{M},\ \boldsymbol{I},\ \boldsymbol{\Sigma})$ 情形.

7. 设 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ 是取自正态总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的 i.i.d 样本, 其中 $\boldsymbol{\Sigma}>0$, $p=2k$, 记 $\boldsymbol{\mu}'=(\boldsymbol{\mu}_{(1)}'\boldsymbol{\mu}_{(2)}')$, $\boldsymbol{\mu}_{(1)}$、$\boldsymbol{\mu}_{(2)}$ 都是 k 维的. 试给出对假设 H_0: $\boldsymbol{\mu}_{(1)}=$

$\boldsymbol{\mu}_{(2)}$ 的检验.

8．设 p 维正态向量 $\boldsymbol{Y} \triangleq (\boldsymbol{Y}'_{(1)}, \cdots, \boldsymbol{Y}'_{(k)})'$，其中 $\boldsymbol{Y}_{(i)}$ 为 p_i 维，$\sum_1^k p_i = p$. 记

$$\operatorname{Cov} \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \boldsymbol{\Sigma}_{1k} \\ \vdots & & \vdots \\ \boldsymbol{\Sigma}_{k1} & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix}$$

试构造检验 $\boldsymbol{Y}_{(1)}, \cdots, \boldsymbol{Y}_{(k)}$ 的独立性($\boldsymbol{\Sigma}_{ij}=0,\ i \neq j$)的检验统计量.

9．设两个独立总体分别有分布 $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ 和 $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，设零假设为

H_0: $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\mu}_0$(已知)

试给出对 H_0 的检验.

10．设 $\boldsymbol{X}, \boldsymbol{Y}$ 分别是 p 维和 q 维正态随机向量，记

$$\operatorname{Cov}\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},\ \boldsymbol{\Sigma}_{11},\ \boldsymbol{\Sigma}_{22} > 0$$

问：为何检验 $\boldsymbol{X}, \boldsymbol{Y}$ 的典型相关系数为零？

11．设 $\boldsymbol{Y} = \boldsymbol{XB} + \varepsilon$ 是多元线性模型.

(i) 给出参数函数 $\operatorname{tr} \boldsymbol{A}'\boldsymbol{B}$ 线性可估的充要条件.

(ii) 当 $\boldsymbol{\varepsilon} \sim N_{np}(\boldsymbol{0}, \boldsymbol{I}, \boldsymbol{\Sigma})$，$\boldsymbol{X} = (\boldsymbol{X}_1 \vdots \boldsymbol{X}_2)$，$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{pmatrix}$，设零假设 H_0: $\boldsymbol{HB}_1 = \boldsymbol{0}$,

试给出检验统计量.

12．(i) 设 $\boldsymbol{Y}_{q\times p} = \begin{pmatrix} \boldsymbol{y}'_1 \\ \vdots \\ \boldsymbol{y}'_q \end{pmatrix}$，$E\boldsymbol{y}_g = \boldsymbol{0}$，$\operatorname{Cov}(\boldsymbol{y}_g, \boldsymbol{y}_h) = \delta_{gh}\boldsymbol{\Sigma}_g$，$\boldsymbol{U}$ 是 q 阶正交阵，$\boldsymbol{U} = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_q)$，其中 $\boldsymbol{u}_q = \frac{1}{\sqrt{q}} \boldsymbol{1}_q$，令 $\boldsymbol{Z} = \boldsymbol{UY}$

证明：$\boldsymbol{Z}$ 的行互不相关 $\Leftrightarrow \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q$.

(ii) 设 $\boldsymbol{X}_\alpha^{(g)}(\alpha = 1, \cdots, n)$ 是取自 $N_p(\boldsymbol{\mu}^{(g)}, \boldsymbol{\Sigma}_g)(g = 1\cdots, q)$ 的独立随机样本，使用(i)中结果构造假设

$$H: \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_q$$

的检验.

13．欲求模型(7.50)中 $\boldsymbol{B}$ 的极小方差线性无偏估计 MVLUE，可由相应的模型(7.52)中，对一元线性模型参数向量 $\boldsymbol{B}$，求 LS 估计而得到，为什么？(看第六章习题 15).

14．试证两个同协方差阵的总体的马氏距离

$$d^2(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

是在非退化线性变化下不变的.

15. 设 $\boldsymbol{X}=(X_1, \cdots, X_p)$, $\boldsymbol{Y}=(Y_1, \cdots, Y_q)$, 是两个随机向量, 其联合分布协方差阵为

$$\operatorname{Cov}\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix}=\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \ \boldsymbol{\Sigma}_{11}>0.$$

(i) 求 $\boldsymbol{b}\in R^p$ 使得

$$\sum_{i=1}^{q} D(Y_i)-\frac{[\operatorname{Cov}(Y_i, \boldsymbol{b}'\boldsymbol{X})]^2}{D(\boldsymbol{b}'\boldsymbol{X})} \quad (*)$$

达极小.

(ii) 求 $\boldsymbol{b}\in R^p$ 使得

$$\sum_{i=1}^{p}\left\{D(X_i)-\frac{[\operatorname{Cov}(X_i, \boldsymbol{b}'\boldsymbol{X})]^2}{D(\boldsymbol{b}'\boldsymbol{X})}\right\}W_i^2,$$

达极小, 其中 W_i^2 是给定的数.

16. 设 $\boldsymbol{y}=\boldsymbol{Z}+\boldsymbol{\varepsilon}$, $E\boldsymbol{Z}=E\boldsymbol{\varepsilon}=0$, $\operatorname{Cov}(\boldsymbol{Z})=\boldsymbol{G}$, $\operatorname{Cov}\boldsymbol{\varepsilon}=\sigma^2\boldsymbol{I}_p$, $\operatorname{Cov}(\boldsymbol{Z}, \boldsymbol{\varepsilon})=0$,

(1) 求 $\boldsymbol{C}\in R^p$ 使 $D(\boldsymbol{C}'\boldsymbol{y})=1$, 且极小化 $D(\boldsymbol{C}'\boldsymbol{\varepsilon})$.

(2) 设 $D(y_i^2)=1$, $i=1, \cdots, p$. 求 $\boldsymbol{C}\in R^p$ 使 $D(\boldsymbol{C}'\boldsymbol{y})=1$ 且极大化 $\sum_{i=1}^{p}\rho^2(y_i, \boldsymbol{C}'\boldsymbol{y})$.

(3) 上述结果与主成分有何联系?

17. 为寻求正交变换 Γ 使(7.82)中的 S 达极大值, 可令

$$\Gamma=\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

试求出 φ 为极大值点时应满足的关系式.

附表1 正态分布函数

本表列出了服从正态分布 $N(0,\ 1)$ 的随机变量 X 的分布函数

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

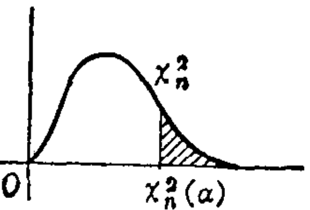
的值.

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.500000	1.05	0.853141	2.10	0.982136
0.05	0.519939	1.10	0.864334	2.15	0.984222
0.10	0.539828	1.15	0.874928	2.20	0.986097
0.15	0.559618	1.20	0.884930	2.25	0.987776
0.20	0.579260	1.25	0.894350	2.30	0.989276
0.25	0.589706	1.30	0.903200	2.35	0.990613
0.30	0.617911	1.35	0.911492	2.40	0.991802
0.35	0.636831	1.40	0.919243	2.45	0.992857
0.40	0.655422	1.45	0.926471	2.50	0.993790
0.45	0.673645	1.50	0.933193	2.55	0.994614
0.50	0.691463	1.55	0.939429	2.60	0.995339
0.55	0.708840	1.60	0.945201	2.65	0.995975
0.60	0.725747	1.65	0.950528	2.70	0.996533
0.65	0.742154	1.70	0.955434	2.75	0.997020
0.70	0.758036	1.75	0.959941	2.80	0.997445
0.75	0.773373	1.80	0.964070	2.85	0.997814
0.80	0.788145	1.85	0.967843	2.90	0.998134
0.85	0.802338	1.90	0.971283	2.95	0.998411
0.90	0.815940	1.95	0.974412	3.00	0.998650
0.95	0.828944	2.00	0.977250		
1.00	0.841345	2.05	0.979818		

附表 2　χ^2-分布

本表对自由度 n 的 χ^2-分布给出上侧分位数 $\chi_n^2(\alpha)$ 表

$$P(\chi_n^2 > \chi_n^2(\alpha)) = \alpha$$



n	α=0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209

（续表）

n	α=0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.669	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

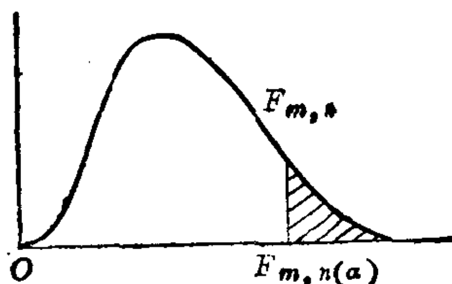
附表 3

本表对 F-分布，给出了关于 $\alpha=0.05$ 的上侧分位数 $F_{m,n}(\alpha)$（上面的数字）和关于 $\alpha=0.01$ 的上侧分位数 $F_{m,n}(\alpha)$（下面的数字）。

$$P(F_{m,n}>F_{m,n}(\alpha))=\alpha$$

分母的自由度	分	子	的								
n	1	2	3	4	5	6	7	8	9	10	11
1	161 4052	200 4999	216 5403	225 5625	230 5764	234 5859	237 5928	239 5981	241 6022	242 6056	243 6082
2	18.51 98.49	19.00 99.01	19.16 99.17	19.25 99.25	19.30 99.30	19.33 99.33	19.36 99.34	19.37 99.36	19.38 99.38	19.39 99.40	19.40 99.41
3	10.13 34.12	9.55 30.81	9.28 29.46	9.12 28.71	9.01 28.24	8.94 27.91	8.88 27.67	8.84 27.49	8.81 27.34	8.78 27.23	8.76 27.13
4	7.71 21.20	6.94 18.00	6.59 16.69	6.39 15.98	6.26 15.52	6.16 15.21	6.09 14.98	6.04 14.80	6.00 14.66	5.96 14.54	5.93 14.45
5	6.61 16.26	5.79 13.27	5.41 12.06	5.19 11.39	5.05 10.97	4.95 10.67	4.88 10.45	4.82 10.27	4.78 10.15	4.74 10.05	4.70 9.96
6	5.99 13.74	5.14 10.92	4.76 9.78	4.53 9.15	4.39 8.75	4.28 8.47	4.21 8.26	4.15 8.10	4.10 7.98	4.06 7.87	4.03 7.79
7	5.59 12.25	4.74 9.55	4.35 8.45	4.12 7.85	3.97 7.46	3.87 7.19	3.79 7.00	3.73 6.84	3.68 6.71	3.63 6.62	3.60 6.54
8	5.32 11.26	4.46 8.65	4.07 7.59	3.84 7.01	3.69 6.63	3.58 6.37	3.50 6.19	3.44 6.03	3.39 5.91	3.34 5.82	3.31 5.74
9	5.12 10.56	4.26 8.02	3.86 6.99	3.63 6.42	3.48 6.06	3.37 5.80	3.29 5.62	3.23 5.47	3.18 5.35	3.13 5.26	3.10 5.18
10	4.96 10.04	4.10 7.56	3.71 6.55	3.48 5.99	3.33 5.64	3.22 5.39	3.14 5.21	3.07 5.06	3.02 4.95	2.97 4.85	2.94 4.78
11	4.84 9.65	3.98 7.20	3.59 6.22	3.36 5.67	3.20 5.32	3.09 5.07	3.01 4.88	2.95 4.74	2.90 4.63	2.96 4.54	2.82 4.46
12	4.75 9.33	3.88 6.93	3.49 5.95	3.26 5.41	3.11 5.06	3.00 4.82	2.92 4.65	2.85 4.50	2.80 4.39	2.76 4.30	2.72 4.22
13	4.67 9.07	3.80 6.70	3.41 5.74	3.18 5.20	3.02 4.86	2.92 4.62	2.84 4.44	2.77 4.30	2.72 4.19	2.67 4.10	2.63 4.02
14	4.60 8.86	3.74 6.51	3.34 5.56	3.11 5.03	2.96 4.69	2.85 4.46	2.77 4.28	2.70 4.14	2.65 4.03	2.60 3.94	2.56 3.86
15	4.54 8.68	3.68 6.36	3.29 5.42	3.06 4.89	2.90 4.56	2.79 4.32	2.70 4.14	2.64 4.00	2.59 3.89	2.55 3.80	2.51 3.73
16	4.49 8.53	3.63 6.23	3.24 5.29	3.01 4.77	2.85 4.44	2.74 4.20	2.66 4.03	2.59 3.89	2.54 3.78	2.49 3.69	2.45 3.61

F-分布



自由度 m												
12	14	16	20	24	30	40	50	75	100	200	500	∞
244 6106	245 6142	246 6169	248 6208	249 6234	250 6258	251 6286	252 6302	253 6323	253 6334	254 6352	254 6361	254 6366
19.41 99.42	19.42 99.43	19.43 99.44	19.44 99.45	19.45 99.46	19.46 99.47	19.47 99.48	19.47 99.48	19.48 99.49	19.49 99.49	19.49 99.49	19.50 99.50	19.50 99.50
8.74 27.05	8.71 26.92	8.69 26.83	8.66 26.69	8.64 26.60	8.62 26.50	8.60 26.41	8.58 26.30	8.57 26.27	8.56 26.23	8.54 26.18	8.54 26.14	8.53 26.12
5.91 14.37	5.87 14.24	5.84 14.15	5.80 14.02	5.77 13.93	5.74 13.83	5.71 13.74	5.70 13.69	5.68 13.61	5.66 13.57	5.65 13.52	5.64 13.48	5.63 13.46
4.68 9.89	4.64 9.77	4.60 9.68	4.56 9.55	4.53 9.47	4.50 9.38	4.46 9.29	4.44 9.24	4.42 9.17	4.40 9.13	4.38 9.07	4.37 9.04	4.36 9.02
4.00 7.72	3.96 7.60	3.92 7.52	3.87 7.39	3.84 7.31	3.81 7.23	3.77 7.14	3.75 7.09	3.72 7.02	3.71 6.99	3.69 6.94	3.68 6.90	3.67 6.88
3.57 6.47	3.52 6.35	3.49 6.27	3.44 6.15	3.41 6.07	3.38 5.98	3.34 5.90	3.32 5.85	3.29 5.78	3.28 5.75	3.25 5.70	3.24 5.67	3.23 5.65
3.28 5.67	3.23 5.56	3.20 5.48	3.15 5.36	3.12 5.28	3.08 5.20	3.05 5.11	3.03 5.06	3.00 5.00	2.98 4.96	2.96 4.91	2.94 4.88	2.93 4.86
3.07 5.11	3.02 5.00	2.98 4.92	2.93 4.80	2.90 4.73	2.86 4.64	2.82 4.56	2.80 4.51	2.77 4.45	2.76 4.41	2.73 4.36	2.72 4.33	2.71 4.31
2.91 4.71	2.86 4.60	2.82 4.52	2.77 4.41	2.74 4.33	2.70 4.25	2.67 4.17	2.64 4.12	2.61 4.05	2.59 4.01	2.56 3.96	2.55 3.93	2.54 3.91
2.79 4.40	2.74 4.29	2.70 4.21	2.65 4.10	2.61 4.02	2.57 3.94	2.53 3.86	2.50 3.80	2.47 3.74	2.45 3.70	2.42 3.66	2.41 3.62	2.40 3.60
2.69 4.16	2.64 4.05	2.60 3.98	2.54 3.86	2.50 3.78	2.46 3.70	2.42 3.61	2.40 3.56	2.36 3.49	2.35 3.46	2.32 3.41	2.31 3.38	2.30 3.36
2.60 3.96	2.55 3.85	2.51 3.78	2.46 3.67	2.42 3.59	2.38 3.51	2.34 3.42	2.32 3.37	2.28 3.30	2.26 3.27	2.24 3.21	2.22 3.18	2.21 3.16
2.53 3.80	2.48 3.70	2.44 3.62	2.39 3.51	2.35 3.43	2.31 3.34	2.27 3.26	2.24 3.21	2.21 3.14	2.19 3.11	2.16 3.06	2.14 3.02	2.13 3.00
2.48 3.67	2.43 3.56	2.39 3.48	2.33 3.36	2.29 3.29	2.25 3.20	2.21 3.12	2.18 3.07	2.15 3.00	2.12 2.97	2.10 2.29	2.08 2.89	2.07 2.87
2.42 3.55	2.37 3.45	2.33 3.37	2.28 3.25	2.24 3.18	2.20 3.10	2.16 3.01	2.13 2.96	2.09 2.89	2.07 2.86	2.04 2.80	2.02 2.77	2.01 2.75

分母的自由度	分子的										
n	1	2	3	4	5	6	7	8	9	10	11
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05
26	4.22	3.37	2.89	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14
	7.60	5.52	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08
	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06
	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.78
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05
	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75

（续表）

自由度 m												
12	14	16	20	24	30	40	50	75	100	200	500	∞
2.38 3.45	2.35 3.35	2.29 3.27	2.23 3.16	2.19 3.08	2.15 3.00	2.11 2.92	2.08 2.86	2.04 2.79	2.02 2.76	1.99 2.70	1.97 2.67	1.96 2.65
2.34 3.37	2.29 3.27	2.25 3.19	2.19 3.07	2.15 3.00	2.11 2.91	2.07 2.83	2.04 2.78	2.00 2.71	1.98 2.68	1.95 2.62	1.93 2.59	1.92 2.57
2.31 3.30	2.26 3.19	2.21 3.12	2.15 3.00	2.11 2.92	2.07 2.84	2.02 2.76	2.00 2.70	1.96 2.63	1.94 2.60	1.91 2.54	1.90 2.51	1.88 2.49
2.28 3.23	2.23 3.13	2.18 3.05	2.12 2.94	2.08 2.86	2.04 2.77	1.99 2.69	1.96 2.63	1.92 2.56	1.90 2.53	1.87 2.47	1.85 2.44	1.84 2.42
2.25 3.17	2.20 3.07	2.15 2.99	2.09 2.88	2.05 2.80	2.00 2.72	1.96 2.63	1.93 2.58	1.89 2.51	1.87 2.47	1.84 2.42	1.82 2.38	1.81 2.36
2.23 3.12	2.18 3.02	2.13 2.94	2.07 2.83	2.03 2.75	1.98 2.67	1.93 2.58	1.91 2.53	1.87 2.46	1.84 2.42	1.81 2.37	1.80 2.33	1.78 2.31
2.20 3.07	2.14 2.97	2.10 2.89	2.04 2.78	2.00 2.70	1.96 2.62	1.91 2.53	1.88 2.48	1.84 2.41	1.82 2.37	1.79 2.32	1.77 2.28	1.76 2.26
2.18 3.03	2.13 2.93	2.09 2.85	2.02 2.74	1.98 2.66	1.94 2.58	1.89 2.49	1.86 2.44	1.82 2.36	1.80 2.33	1.76 2.27	1.74 2.23	1.73 2.21
2.16 2.99	2.11 2.89	2.06 2.81	2.00 2.70	1.96 2.62	1.92 2.54	1.87 2.45	1.84 2.40	1.80 2.32	1.77 2.29	1.74 2.23	1.72 2.19	1.71 2.17
2.15 2.96	2.10 2.86	2.05 2.77	1.99 2.66	1.95 2.58	1.90 2.50	1.85 2.41	1.82 2.36	1.78 2.28	1.76 2.25	1.72 2.19	1.70 2.15	1.69 2.13
2.13 2.93	2.08 2.83	2.03 2.74	1.97 2.63	1.93 2.55	1.88 2.47	1.84 2.38	1.80 2.33	1.76 2.25	1.74 2.21	1.71 2.16	1.68 2.12	1.67 2.10
2.12 2.90	2.06 2.80	2.02 2.71	1.96 2.60	1.91 2.52	1.87 2.44	1.81 2.35	1.78 2.30	1.75 2.22	1.72 2.18	1.69 2.13	1.67 2.09	1.65 2.06
2.10 2.87	2.05 2.77	2.00 2.68	1.94 2.57	1.90 2.49	1.85 2.41	1.80 2.32	1.77 2.27	1.73 2.19	1.71 2.15	1.68 2.10	1.65 2.06	1.64 2.03
2.09 2.84	2.04 2.74	1.99 2.66	1.93 2.55	1.89 2.47	1.84 2.38	1.79 2.29	1.76 2.24	1.72 2.16	1.69 2.13	1.66 2.07	1.64 2.03	1.62 2.01
2.07 2.80	2.02 2.70	1.97 2.62	1.91 2.51	1.86 2.42	1.82 2.34	1.76 2.25	1.74 2.20	1.69 2.12	1.67 2.08	1.64 2.02	1.61 1.98	1.59 1.96
2.05 2.76	2.00 2.66	1.95 2.58	1.89 2.47	1.84 2.38	1.80 2.30	1.74 2.21	1.71 2.15	1.67 2.08	1.64 2.04	1.61 1.98	1.59 1.94	1.57 1.91
2.03 2.72	1.89 2.62	1.93 2.54	1.87 2.43	1.82 2.35	1.78 2.26	1.72 2.17	1.69 2.12	1.65 2.04	1.62 2.00	1.59 1.94	1.56 1.90	1.55 1.87
2.02 2.69	1.96 2.59	1.92 2.51	1.85 2.30	1.80 2.32	1.76 2.22	1.71 2.14	1.67 2.08	1.63 2.00	1.60 1.97	1.57 1.90	1.54 1.86	1.53 1.84

分母的自由度	分 子 的										
n	1	2	3	4	5	6	7	8	9	10	11
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02
	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01
	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00
	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99
	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98
	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97
	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95
	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94
	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.79	2.70	2.61	2.54
70	3.98	3.13	2.74	2.50	2.35	2.32	2.14	2.07	2.01	1.97	1.93
	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91
	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88
	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86
	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.65	2.56	2.47	2.40
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85
	6.81	4.75	3.91	3.44	3.13	2.92	2.76	2.62	2.53	2.44	2.37
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83
	6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81
	6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80
	6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79
	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24

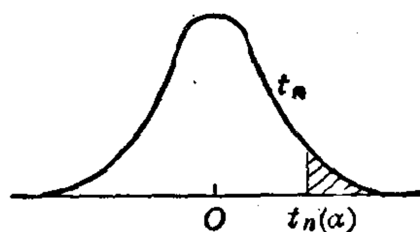
（续表）

自由度 m												
12	14	16	20	24	30	40	50	75	100	200	500	∞
2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
2.66	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84	1.81
1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
2.64	2.54	2.46	2.35	2.26	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78
1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
2.62	2.52	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75
1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
2.60	2.50	2.42	2.40	2.22	2.13	2.04	1.98	1.90	1.86	1.80	1.76	1.72
1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
2.58	2.48	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70
1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
2.56	2.46	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.76	1.71	1.68
1.9[illegible]	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
2.53	2.43	2.35	2.23	2.15	2.06	1.96	1.90	1.82	1.78	1.71	1.66	1.64
1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
2.50	2.40	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
2.47	2.30	2.27	2.18	2.09	2.00	1.90	1.84	1.76	1.71	1.64	1.60	1.56
1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
2.45	2.35	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.63	1.56	1.53
1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
2.41	2.32	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49
1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
2.36	2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
2.33	2.23	2.15	2.03	1.94	1.85	1.75	1.68	1.59	1.54	1.46	1.40	1.37
1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
2.30	2.20	2.12	2.00	1.91	1.83	1.72	1.66	1.56	1.51	1.43	1.37	1.33
1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
2.28	2.17	2.09	1.97	1.88	1.79	1.69	1.62	1.53	1.48	1.39	1.33	1.28
1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
2.23	2.12	2.04	1.92	1.84	1.74	1.64	1.57	1.47	1.42	1.32	1.24	1.19
1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
2.20	2.09	2.01	1.89	1.81	1.71	1.61	1.54	1.44	1.38	1.28	1.19	1.11
1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
2.18	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00

附表4 t-分布

本表对 t-分布给出上侧分位数 $(t_n(\alpha))$ 表，其中 n 是自由度。

$$P(t_n > t_n(\alpha)) = \alpha$$



n \ α	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576